# CROSS-LINGUAL TEXT ANNOTATION

Tadej Štajner

**Doctoral Dissertation**
**Jožef Stefan International Postgraduate School**
**Ljubljana, Slovenia**

**Supervisor:** Prof. Dunja Mladenić, Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

**Evaluation Board:**
Prof. Tomaž Erjavec, Chair, Jožef Stefan Institute, Ljubljana, Slovenia and Jožef Stefan International Postgraduate School, Slovenia
Prof. Marko Tadić, Member, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Prof. Marko Robnik-Šikonja, Member, Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Tadej Štajner

# CROSS-LINGUAL TEXT ANNOTATION

**Doctoral Dissertation**

# MEDJEZIKOVNO OZNAČEVANJE BESEDIL

**Doktorska disertacija**

**Supervisor:** Prof. Dunja Mladenić

Ljubljana, Slovenia, May 2018

# Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (ICT-216886-NoE), ACTIVE (IST-2008-215040), XLike (ICT-STREP-288342), RENDER (ICT-257790-STREP), XLime (FP7-ICT-611346) and LT-Web (FP7-ICT-287815-CSA).

# Abstract

This dissertation discusses the challenges of integrating a new language in an intelligent system. A major issue in natural language processing is mixing multi-lingual content within the same system. We introduce the notion of cross-linguality, a strategy of handling multilingual content without machine translation pipelines. We demonstrate the possibilities and weaknesses of individual proposed solutions and what kind of abstractions they let us use when dealing with multilingual content.

The main motivation for the dissertation is learning how can we use knowledge bases in improving access to multilingual information, and how can we better understand user-generated content.

First, we focus on the different roles of knowledge bases in multilingual information access. We tackle the challenge of using multilingual knowledge bases as comparable corpora to solve cross-lingual tasks, such as bilingual dictionary induction, similarity estimation and vector space translation, tasks that are central to multilingual information access. We propose a method using word embeddings and kernel approximation to train scalable non-linear transformations. We demonstrate that this novel method works better on a majority of evaluated language pairs. We also obtain results in posing vector space translation as a sparse signal reconstruction problem, treating the cross-linguality as a hypothetical noisy channel, where our task is to reconstruct the signal in the target language from measurements in our source language.

As a second use case of knowledge bases, we present a novel method for named entity disambiguation based on a notion of relatedness among entities. We present complementary approaches of representing relatedness and show their respective contributions to the quality of the result. We also discuss the cross-lingual version of this problem, along with the challenges when developing a named entity recognition system for Slovene.

We also discuss the notion of language in social media as a genre of user-generated language that does not conform to assumptions in typical natural language processing systems. We present a novel method for summarizing streams of interesting responses to news in social media in the form of a sampling algorithm. We show its performance on real-life scenarios, as well as use it to demonstrate interesting responses, shedding light on the subjectivity around the selection process and the perception of interestingness.

As another use case of user generated content, we present a novel method for sentiment analysis that uses meta-modeling and integration of background knowledge. We show its performance across domains and languages, showing that it can provide good performance even on small training sets.

# Povzetek

Disertacija obravnava izzive, ki se pojavljajo ob integraciji novega jezika v inteligentni sistem. Eden večjih izzivov pri procesiranju naravnega jezika je uporaba več jezikov znotraj istega sistema. V ta namen definiramo koncept medjezičnosti kot strategijo za reševanje problemov, ki zajemajo večjezično vsebino, brez da bi se posluževali celotnih sistemov za strojno prevajanje. Disertacija prikaže in ovrednoti prednosti in slabosti posameznih predstavljenih komponent ter opiše različne abstrakcije, ki so nam na voljo pri procesiranju večjezičnega besedila. Temelj medjezičnosti je namreč zmožnost, da lahko obravnavamo večjezično besedilo na skupni predstavitvi.

Glavna motivacija te disertacije je ugotoviti, kako lahko baze znanja pripomorejo k dostopu do večjezičnih informacij in razumevanju uporabniško ustvarjene vsebine.

Najprej se osredotočimo na različne vloge baz znanja pri dostopu do večjezičnih informacij. Osredotočimo se na temeljni izziv medjezičnosti, grajenje prevajalnih modelov z uporabo šibko poravnanih dvojezičnih korpusov, kar je uporabno v primerih, kjer ni možna uporaba paralelnih korpusov. Modele ovrednotimo na večih nalogah: prevajanje v vektorskem prostoru, ugotavljanja medsebojne podobnosti dokumentov v različnih jezikih in sestavljanje dvojezičnih slovarjev. Predstavimo metodo na osnovi vložitve besed in nelinearne regresije, kjer se poslužujemo učinkovitih metod za učenje jedrne regresije na podlagi približkov. Druga predstavljena metoda se osredotoča na interpretacijo problema s stališča rekonstrukcije redkega signala, kjer prevajanje vidimo kot komunikacijski kanal s šumom, naša naloga pa je rekonstrukcija vektorske oblike besed iz izvornega v ciljni jezik.

Kot drugi način uporabe baz znanja predstavimo tudi metodo za razločevanje imenskih entitet v besedilih na podlagi povezanosti med entitetami. Predstavimo več načinov obravnavanja povezanosti, ki se izkažejo kot komplementarni. Med temi predstavimo novo metodo za izračun povezanosti med parom entitet na podlagi statističnega uteževanja medsebojnih povezav, kjer jo primerjamo z uporabo statistike sopojavitev entitet ter uporabo podobnosti besedil kot metrike podobnosti entitet. Pokažemo, da najboljši rezultat prinese uporaba vseh treh metrik.

Disertacija obravnava tudi jezik v družbenih medijih kot specifičen žanr uporabniško ustvarjene vsebine, kjer mnoge predpostavke, ki se tipično pojavljajo pri procesiranju naravnega jezika, ne veljajo več. V ta namen uporabimo drugačne predstavitve jezika. Predstavimo metodo za povzemanje komentarjev v družabnih medijih na novice, kjer želimo prikazati le najzanimivejše odzive na novico. Predstavimo nov algoritem za povzemanje, ki ima tudi teoretične garancije za kakovost svojega delovanja. Delovanje metode primerjamo z več primerljivimi metodami in s pomočjo dveh različnih evalvacij prikažemo naše izboljšave. Delovanje prikažemo na resničnih podatkih, s čimer tudi obravnavamo povezanost koncepta zanimivosti z drugimi lastnostmi besedila.

Predstavimo tudi novo metodo za analizo naklonjenosti, ki se lahko prilagodi različnim žanrom besedila s pomočjo uporabe zunanjega znanja ter nebesednih značilk s poudarkom na visoki natančnosti že pri majhnih učnih množicah.

# Contents

**Abstract** **vii**

**Povzetek** **ix**
**List of Figures** **xv**

**List of Tables** **xvii**

**List of Algorithms** **xix**

**Abbreviations** **xxi**

**1 Introduction** **1**
1.1 Scientific Contributions . . . . . 3

**2 Related work** **5**
2.1 Leveraging Knowledge Bases for Cross-Linguality . . . . . 5
2.1.1 Document-level approaches to cross-linguality . . . . . 7
2.1.2 Named entity disambiguation . . . . . 8
2.2 Language in Social Media . . . . . 9
2.2.1 Behavior in social media . . . . . 9
2.2.2 Sentiment analysis across domains and languages . . . . . 10

**3 Leveraging Knowledge Bases for Cross-Linguality** **11**
3.1 Introduction . . . . . 11
3.1.1 Use cases . . . . . 11
3.1.2 Multi-lingual language resources . . . . . 12
3.1.3 Dataset description . . . . . 13
3.1.4 Summary of scientific contributions . . . . . 13
3.2 Cross-Lingual Document Similarity Estimation and Dictionary Induction with Comparable Corpora . . . . . 15
3.2.1 Introduction . . . . . 15
3.2.1.1 Word embeddings . . . . . 16
3.2.2 Proposed approach . . . . . 18
3.2.2.1 Monolingual embeddings with a kernel approximation mapping . . . . . 18
3.2.2.2 Monolingual embeddings with an ExtraTrees mapping . . . 19
3.2.2.3 Monolingual embeddings with a cross-lingual CCA mapping 19
3.2.3 Baselines and related approaches . . . . . 21
3.2.3.1 Cross-lingual Latent Semantic Indexing . . . . . 21
3.2.3.2 Low-rank Canonical Correlation Analysis . . . . . 22
3.2.3.3 Regression CCA . . . . . 24

3.2.4 Experiments . . . 25
3.2.4.1 Evaluation dataset . . . 25
3.2.4.2 Experiments on bilingual dictionary induction . . . 26
3.2.4.3 Experiments on cross-lingual similarity estimation . . . 29
3.2.5 Discussion . . . 30
3.3 Sparsity in Cross-Lingual Translation . . . 31
3.3.1 Problem formulation . . . 32
3.3.2 Elastic CCA . . . 32
3.3.3 Experiments . . . 33
3.3.4 Comparison across language pairs . . . 35
3.3.5 Summary . . . 36
3.4 Named Entity Disambiguation . . . 37
3.4.1 Problem Definition . . . 37
3.4.2 Using entity relatedness as a coherence estimate . . . 38
3.4.2.1 Independent entity disambiguation baseline . . . 39
3.4.2.2 Collective resolution with relatedness . . . 39
3.4.2.3 Using semantic relations from the knowledge base . . . 40
3.4.2.4 Using content similarity . . . 42
3.4.2.5 Using co-occurrences from an annotated corpus . . . 42
3.4.2.6 Combining relatedness estimations . . . 43
3.4.2.7 Evaluation . . . 43
3.4.2.8 Discussion . . . 46
3.4.3 Cross-lingual named entity disambiguation . . . 46
3.4.3.1 Context similarity . . . 46
3.4.3.2 Evaluation . . . 48
3.4.4 Named entity extraction for Slovene . . . 48
3.4.5 Discussion . . . 49
3.4.6 Applications in internationalization . . . 50
3.4.6.1 Data and interaction model . . . 52
3.4.6.2 Example . . . 53
3.4.6.3 Discussion . . . 53

**4 Language in Social Media 55**
4.1 Automatic Selection of Interesting Social Media Responses to News . . . 57
4.1.1 Proposed method . . . 58
4.1.1.1 Sampling algorithm . . . 59
4.1.1.2 Message scoring and message set selection . . . 59
4.1.1.3 Comparison with the state of the art . . . 61
4.1.2 Experiments . . . 63
4.1.2.1 Dataset . . . 63
4.1.2.2 Gold standard collection . . . 64
4.1.2.3 Experimental evaluation . . . 65
4.1.2.4 Preference judgment collection . . . 67
4.1.2.5 Preference judgment analysis . . . 68
4.1.2.6 Method comparison . . . 69
4.1.3 Feature analysis . . . 70
4.1.3.1 Individual message features . . . 70
4.1.3.2 Learning to rank message sets . . . 73
4.2 Informal Multilingual Multi-Domain Sentiment Analysis . . . 76
4.2.1 Sentiment lexicons . . . 76
4.2.2 Feature construction . . . 77

4.2.3 Models . . . 78
4.2.3.1 Feature combination . . . 78
4.2.4 Experiments . . . 80
4.2.5 Discussion . . . 83

**5 Conclusions** **85**
5.1 Leveraging Knowledge Bases for Cross-Linguality . . . 85
5.2 Named Entity Disambiguation . . . 86
5.3 Automatic Selection of Interesting Social Media Responses to News . . . 87
5.4 Sentiment Analysis in Informal Multilingual Multi-Domain Content . . . 88
5.5 Summary . . . 88

**References** **91**

**Bibliography** **103**

**Biography** **105**

# List of Figures

Figure 3.1: Integration of word embedding representations and various cross-lingual linear mappings. . . . 17
Figure 3.2: Integration of kernel approximation as a non-linearity in the cross-lingual translation pipeline. . . . 19
Figure 3.3: Multivariate ExtraTree estimators as a non-linear cross-lingual mapping. 20
Figure 3.4: Monolingual embeddings with a CCA cross-lingual transform (CCAT-SKIP and CCAT-CBOW). . . . 20
Figure 3.5: Architecture of Cross-lingual LSI, showing how the target $y$ is obtained from the input $x$ . . . 23
Figure 3.6: Architecture of the low rank CCA bidirectional translation pipeline. . . 23
Figure 3.7: Illustration of Regression CCA, showing the direct linear transformation expression. . . . 24
Figure 3.8: Visualization of results for every language pair given their Mean Reciprocal Rank score. The vertical axis is the source language, and the horizontal axis represents the target language. Darker tones correspond to better performance. . . . 28
Figure 3.9: Visualization of results for every language pair given their $R^2$ coefficients. The luminance is an exponential transform of $R^2$, darker tones corresponding to better performance. . . . 35
Figure 3.10: Using different semantic relations as a relatedness measure. . . . 40
Figure 3.11: Using content similarity as a relatedness measure. The green dotted lines represent selected entities. . . . 42
Figure 3.12: Using mutual information form entity co-occurrences as relatedness. . . 42
Figure 3.13: Comparison of entity disambiguation quality of different relatedness heuristics. . . . 45
Figure 3.14: Architectures for cross-lingual named entity disambiguation . . . 47

Figure 4.1: The social media message selection problem: selecting a small subset of messages sent in response to a news article. . . . 58
Figure 4.2: Proposed approach . . . 60
Figure 4.3: Importances of features . . . 72
Figure 4.4: Importances of features . . . 75
Figure 4.5: Feature concatenation diagram. . . . 78
Figure 4.6: Separate bag-of-words model, denoted as "Words and features". . . . 79
Figure 4.7: Separate model for every feature set, aggregated by voting. . . . 79
Figure 4.8: Meta-classifier, using class probabilities from the inner classifier predictions as its features. . . . 79
Figure 4.9: Sentiment $F_1$ scores with various sentiment lexicons for English. . . . 81
Figure 4.10: Sentiment $F_1$ scores with various sentiment lexicons for Spanish. . . . 82
Figure 4.11: $F_1$ scores with various feature combination approaches across both languages and two learning approaches. . . . 82

Figure 4.12: Using various feature sets on English datasets, using W+F-SVM as the model. . . . . 83
Figure 4.13: Sentiment $F_1$ scores on Spanish datasets, using W+F-SVM as the model. 84

# List of Tables

Table 3.1: Number of aligned documents among selected languages of Wikipedia. . 13
Table 3.2: An overview of all methods used in this evaluation, showing the differences in representation learning and cross-lingual mapping components, where applicable. The methods' segments represent the hypothesis that they are designed to test. . . . . 25
Table 3.3: Precision-at-10 of different method on different language pairs on dictionary induction. . . . . 26
Table 3.4: Precision-at-10 of dictionary induction of different approaches on different language pairs, measuring with Microsoft Translator Gold Standard. 28
Table 3.5: Spearman and Pearsons's correlation coefficients between the predictions and ground truth similarity judgments. . . . . 30
Table 3.6: Effect of thresholding strategies on performance. . . . . 34
Table 3.7: Average $R^2$ and MRR results across all language pairs for both thresholded and non-thresholded variant of models, as well as the natively sparse output of ElasticCCA. . . . . 34
Table 3.8: Performance of respective methods on entity resolution. . . . . 44
Table 3.9: Recall at two levels of precision: 90% and 95% precision. . . . . 45
Table 3.10: Performance of approaches on cross-lingual named entity disambiguation. 48
Table 3.11: Comparing results on named entity extraction, measuring $F_1$ scores on multiple entity classes. . . . . 48

Table 4.1: The features used by our methods. Superscripts $s$ and $d$ indicate scoring (used in $r(m)$) and diversity (used in $H_0(S)$) features. . . . . 62
Table 4.2: Comparison of the tweet sets generated by the methods in terms of their informativeness, opinionatedness and interestingness as measured by ROUGE-2 and $F_1$ at $k = 10$. The asterisk marks the cases where the measurements were significantly higher than all of the three baselines, having a T-test p-value below 0.05. . . . . 66
Table 4.3: Comparison of tweet selection methods based on annotator preference votes, along with inter-annotator agreement via Fleiss' kappa $\kappa_f$. Scores from 0.00 to 0.40 indicate slight agreement, while 0.40 to 0.75 indicate good agreement. . . . . 69
Table 4.4: Distribution of preference reasons for the preferred method in a given method pair. . . . . 71
Table 4.5: Performance of pairwise ranking prediction on various feature sets. The bold text indicates that the result is significantly higher than the RankSVM-$\mu$ baseline. . . . . 74

# List of Algorithms

ALGORITHM 3.1: Training of monolingual embeddings with a kernel approximation mapping model . . . . . 21
ALGORITHM 3.2: Prediction with monolingual embeddings with a kernel approximation mapping model . . . . . 22
ALGORITHM 3.3: Collective entity resolution algorithm. . . . . 40

ALGORITHM 4.1: Greedy iterative sampling algorithm . . . . . 59

# Abbreviations

| | | |
|---|---|---|
| CCA | ... | Canonical Correlation Analysis |
| CCAR | ... | Canonical Correlation Analysis Regression |
| CL-LSI | ... | Cross-lingual Latent Semantic Indexing |
| CRF | ... | Conditional Random Fields |
| IRI | ... | Internationalized Resource Identifier |
| ITS 2.0 | ... | Internationalization Tag Set 2.0 |
| LR-CCA | ... | Low-rank Canonical Correlation Analysis |
| LSI | ... | Latent Semantic Indexing |
| NED | ... | Named Entity Disambiguation |
| NER | ... | Named Entity Recognition |
| NLP | ... | Natural Language Processing |
| SVM | ... | Support Vector Machine |
| URI | ... | Uniform Resource Identifier |

# Chapter 1

# Introduction

More and more of the content on the internet is written in a language other than English. But, not all language tools are available for all languages. For instance, minority languages often tend to lag behind bigger languages. Furthermore, language tools are rarely reusable across different languages.

The goal of this dissertation is to use information contained in unstructured multilingual text data in a language-independent, structured and consistent representation. Text annotations allow us to reason about textual content on a higher level and interact with not just with text, but rather people, places, companies, events, sentiments and relationships. We define *cross-linguality* as a strategy of solving a text annotation problem, without relying on the translation of the input.

We can characterize the problem space within this thesis by three axes:

**Annotation depth** represents the complexity of the annotation. Simpler annotations have the form of a single label per document, such as sentiment or topic classification. Annotations like morpho-syntactic tags or named entities have the form of a sequence annotation. On the deeper extreme, we have structured annotation tasks, such as parsing and relation extraction.

**Domain breadth** represents the strength of the assumptions we can assume about the world within a given task. If the problem definition assumes a small domain or a specific genre, it has a narrower context. A task that needs to work on an open domain has a wider context. For example, classifying sentiment in movie reviews is a domain-specific task, while linking named entities to a general-domain knowledge base assumes any well-formed text. Domain breadth is particularly important for dealing with user-generated content: the experiments in this thesis rely on diverse data sources such as Wikipedia and Twitter.

**Multi-linguality** tells us whether the task is defined within one language, a pair of languages, or a set of languages. With mono-lingual tasks, assumptions take place in a single language. With a multi-lingual task, we can see language as one of the properties of the input.

We summarize the motivation for the work described in this thesis as answering the following questions:

- How can we use knowledge bases in improving access to multilingual information?
- How can we better understand user-generated content?

The text annotation problem domain span over multiple disciplines:

**Machine learning** deals with the mathematical modeling, formalization, and the computational details of an annotation task. At this level, we formalize annotation problems

as prediction tasks. This typically takes place in the form of supervised learning, where a model predicts a value of a target variable given its parameters that were fit on a training dataset. Another relevant form of machine learning is representation learning: discovering a representation of text, better suited for a given task.

**Natural language processing** (NLP) deals with the challenges of obtaining structured information from language. Natural language processing components enable operating on different representations of text. The representation layers range from the morpho-syntactic annotations, lemmas, noun phrases, named entities, up to parse trees and relationship extraction. The core assumption of NLP is that unstructured textual data contains structured information that can be re-constructed using appropriate techniques.

**Knowledge management and engineering** deal with the challenges surrounding structure and semantics of the data and how they relate to the human interaction with it. The capabilities, of knowledge management and engineering components give us the ability to query, reason, manipulate and integrate the data in a well-defined world model. The core assumption is that structured information extracted from textual content is a representation of some coherent world model. It provides the abstraction in the form of a knowledge base, defining a data model for implementing interfaces for human-computer or computer-computer interaction in a language-independent way.

A common denominator in many of the tasks described in this thesis is coming up with a representation of text that is amenable to a multi-lingual setting. The bag-of-words model is commonly used as a baseline feature representation. However, meaning and knowledge cannot be transferred easily when we consider words as independent units. This raises a question on whether we can use a more appropriate representation of content.

Work described in this dissertation uses two different representation approaches besides bag-of-words. The first are embeddings (Bengio, Schwenk, Senécal, *et al.* [1]), a continuous lower-dimensional representation that optimizes for preserving the local context of every word. The second uses explicit semantic representation of a document (Gabrilovich and Markovitch [2]), which describes every document in a vector space defined by its connection to entities of a knowledge base. We call the process of obtaining such a representation entity extraction and resolution.

Named entity extraction is a technique that identifies fragments of text that can be interpreted as named entities of certain types, such as locations, people, organizations, products, events and others. The natural language processing community also treats numeric or monetary expressions as named entities due to the fact that the same methods can be used for extraction, even though they are not 'entities' in the traditional definition of the word.

Using named entity extraction and resolution can provide links from literal terms to concrete concepts, even in ambiguous circumstances. It is inherently a difficult problem: A name can refer to multiple entities, an entity can have multiple names. However, its output allows us to reason and interact with the text on a more meaningful level.

Another important concept in this thesis is using user-generated content to help us solve multilingual problems. Particularly, we explore the challenge of using user-generated comparable corpora to build dictionary induction or similarity estimation models, a hard problem due to the weaker assumptions we can make about the data.

On the applications side, we motivate our efforts with use cases that rely on the better understanding of user-generated content in social media, where we are typically dealing with a wider context and linguistic variety.

On the topic of understanding user-generated data, we observe that social media services are increasingly playing a major role as platforms for people to express and share

their opinions on current events. Responses to real-world events often come in the form of short messages, referring to specific topics, content or information sources. Such messages serve several purposes: to share a particular news article, to express an opinion about the ongoing events, or to add or refute information about the topic or the mentioned article.

On another front, sentiment analysis is a natural language processing task which aims to predict the polarity (usually denoted as positive, negative or neutral) of users publishing sentiment data, in which they express their opinions. The task is traditionally tackled as a classification problem using supervised machine learning techniques. However, this approach requires additional effort in manual labeling of examples and often has difficulties in transferring to other domains. We explored options on how to make the problem of extending to new domains and languages as simple as possible.

Within the dissertation, we first focus on the idea of cross-linguality and propose approaches on problems involving both multi-linguality and different annotation depths, such as bilingual dictionary induction, similarity estimation, and named entity disambiguation. Following that, we focus on the domain of social media, and focus on two tasks involving different domain breadths: summarization of social media responses, and sentiment analysis.

## 1.1 Scientific Contributions

The scientific contributions of this thesis are summarized as:

- A novel method for building bilingual dictionary induction and similarity estimation using comparable corpora. We present several improvements to the baseline models that use different properties of the problem domain, including representation learning to take advantage of monolingual training data, kernel approximation to efficiently learn complex cross-lingual relationships, and sparse reconstruction that allows us to give sparse solutions that correspond better to real-world dictionary induction scenarios.

- A novel method for entity disambiguation in texts that is based on a notion of relatedness among entities. We present several complementary approaches of representing relatedness and show their respective contributions to the quality of the result, both in a monolingual and a cross-lingual setting.

- A novel method for summarizing streams of responses to news in social media. We show the performance of the proposed model on two different experiments, as well as use it to demonstrate interesting responses, shedding light on the subjectivity around the selection process and the perception of interestingness.

- A novel method for sentiment analysis that uses meta-modeling and integration of background knowledge. We show its performance across several domains and languages, showing that it can provide good performance even on small training sets.

We report more specific scientific contributions and quantitative results in the experimental sections of the respective topics.

# Chapter 2

# Related work

In developing multi-lingual natural language processing systems, supporting novel languages is still difficult due to the fact that a large portion of the entire stack needs to handle the details of that language. Ideally, a solution to cross-linguality would be a language-independent semantic representation, such as the Resource Description Framework (RDF) [3].

When faced with a challenge that entails supporting multiple languages, one potential solution is re-implementing the entire pipeline for every additional language, a decision that comes with extra development costs. Another possibility is to use a machine translation system that translates the text so that we can use a pipeline for a language which we already support through a translation model, and processing the translated in a component intended for that target language. Such approaches tend to work sufficiently well for text mining [4], as well as in problems such as cross-lingual entity linking [5] and sentiment analysis [6]. For scenarios, such as cross-lingual spam classification [7], it has already been established that a similar approach using a bag-of-words translation can be sufficient. To summarize, we can solve natural language processing tasks that operate on the bag-of-words level of abstraction using machine translation, especially information retrieval use cases.

Cross-linguality covers the third option, keeping the development scalable, but avoiding introducing a machine translation system. In this chapter, we discuss the different approaches in the literature and use-cases relevant to this thesis.

## 2.1 Leveraging Knowledge Bases for Cross-Linguality

We can divide approaches to cross-lingual use cases into three groups, according to assumptions on the training data.

The first group operates on **word-level** alignments in the form of bilingual dictionaries. For example, recent representation learning approaches based on neural network language models [8] have shown that representing this matrix in a lower-dimensional embedding is computationally efficient and still performs well for the task of translating specific words. The authors pose the problem of translation as a combination of first learning a low-dimensional representation of the aligned corpora, and then calculating the translation matrix that translates one low-dimension. Similarly, [9] produces multilingual word-level alignments based on the intuition that lexico-semantic content should additionally be invariant across languages. Sizable bilingual dictionaries are not available for many language pairs, and the quality of automatic word alignment greatly affects the quality of the embeddings. It is also unclear whether the embedding process provides significant added value beyond the initial word alignments [10]. While a dictionary provides a context-

free approach to translation, they are a special case of a translation matrix. However, a full-blown translation matrix can be costly to represent in a fully materialized fashion.

Approaches in the second group operate on **sentence-level** alignments, common in legal translations and religious texts .The main data point in such datasets is a bilingual pair of sentences, also known as "parallel corpora". This is the main form of training data for the majority of machine translation approaches.

From a performance viewpoint, machine translation trained on sentence-level alignments represents an upper bound for a computer-based system. Another benefit in operating with sentence-level alignments in that we can get a strong baseline already by using sentence identifiers as features when representing individual words (Levy, Søgaard, and Goldberg [11]), a mechanism that is unavailable for document-level alignments.

For the class of tasks that operate sufficiently well on the bag-of-words text representation, such as topic classification, another typical approach is using a bi-lingual dictionary that maps individual words from one language to another [12]. In terms of performance, dictionary mapping gives us a lower bound on performance, given that it is a context-free method that always projects the same word to the same output.

The third group makes a much weaker assumption, **document-level** alignments, and uses comparable texts in different languages (not necessarily translations), such as Wikipedia articles or news reports of the same event. Algorithms in this group try to leverage the large amounts of available data to compensate for the lack of lower-level alignments [13]. One such approach, as proposed in Søgaard, Agić, Alonso, *et al.* [14], describes a count-based model which builds binary word vectors denoting the occurrence of each word in each document pair. Dimensionality reduction is then applied afterwards on the induced sparse vectors. Since the links between documents are known, the model is able to learn cross-lingual correspondences between words and, consequently, bilingual word representations.

Another important axis to characterize cross-lingual representation learning is distinguishing between **parallel** and **comparable** alignments [15]. While parallel corpora require both sides to be exact translations of one another, comparable corpora relax that constraint. Comparable corpora are collections of documents that have certain common features, such as topic, domain, or genre, but do not have an explicit source-target relationship. In the case of Wikipedia, it contains cross-lingual links among documents of the same topic in different languages. Given these links, we work with the assumption that words that have the same meaning in different languages tend to appear in the same lexical contexts. Comparable corpora have the following defining characteristics, when contrasting them with parallel corpora [16]:

- Words have multiple senses per corpus
- Words have multiple translations per corpus
- Translations might not exist in the corpus
- Frequencies of occurrence not comparable
- Positions of occurrence not comparable

The work in this thesis focuses on using **comparable document-level** datasets. Such approaches can be further distinguished as:

**Pseudo-bilingual document-aligned corpora** automatically construct a pseudo-bilingual corpus containing words from both the source and target languages by combining words from aligned documents, such as [13], where both languages are mixed to contribute to the context used to train embeddings.

**Concept-based methods** leverage information about the distribution of latent topics or concepts in document-aligned data to represent words. Such methods [17] are based on the intuition that words in source and target language are similar if they are likely to generate similar words as their top semantic word responses.

**Extensions of sentence-aligned models** extend methods using sentence-aligned parallel data to also work without parallel data. [18] is one such example, describing a model to learn bilingual embeddings without word alignments either from sentence-aligned parallel or label-aligned non-parallel document corpora to support cross-lingual text classification.

For example, we consider the work described in Chapter 3 as an example of the third category. However, the experimental work focusing on the particularities of document-aligned datasets.

### 2.1.1 Document-level approaches to cross-linguality

A relatively simple approach to obtaining a low-dimensional text representation for use in multilingual contexts is Cross-lingual LSI (CL-LSI), a method that has been known to work well for cross-lingual information retrieval [19], leveraging the idea of latent concepts.

Recently, representation learning approaches based on distributed representations of words have shown promising results in various natural language processing tasks. The Skip-gram and Continuous bag-of-words word embedding architectures [20] have been shown to learn semantically-meaningful continuous representations of words from plain text corpora by leveraging the words' neighborhoods as contexts. The model learns to represent a word as a fixed length embedding vector by predicting adjacent words in the document from this vector representation, or predicting the most likely word given a context. These have also been applied to the task of vector space translation [8]. The authors show that we can represent the relationship between two languages in a very compact model.

Another family of methods that is frequently encountered in cross-lingual text analysis scenarios is Canonical correlation analysis (CCA) that addresses the following optimization problem: given two comparable corpora matrices, find a pair of linear mappings that maximize the correlation between variables. The intent of those mappings is to translate examples from both spaces into a common subspace. This computational framework allows us to not use parallel corpora, but also the more easily-available comparable corpora, such as Wikipedia, that has already shown to be a good experimental dataset for cross-lingual information retrieval [21].

CCA is a method of exploring correlations between different representations of some underlying data where these representations exist in two different feature spaces (vector spaces). For example, in case of modeling the semantics of music, data is represented in the form of text labels and audio features. CCA has also been used in applications dealing with multi-language text analysis [22], learning semantic representations between images and text [23]. Similar to the way principal component analysis (PCA) finds informative vectors in one feature space by maximizing the variance of projected data, CCA on the other hand finds pairs of vectors within the two spaces that maximize correlation.

A common application for this transform is cross-lingual information retrieval, which can be done by performing similarity search across a corpus of documents in another language, represented in that common space [24]. While these approaches map directly from one high-dimensional representation to another, it is also possible to combine them with a separate low-rank transform that speeds up the common space operation itself. The intuition behind Low-rank CCA Translation (LR-CCA) [25] is that the cross-lingual mappings can be more compactly represented in a low-dimensional basis instead of the high-dimensional word spaces.

In terms of applications, literature [26] demonstrates cross-lingual named entity recognition by using a combination of two approaches: first, an annotation mapping based on comparable corpora, and second, a cross-lingual embedding mapping that is used instead of re-training the model for another language.

Furthermore, knowledge base population, a task of finding additional assertion triplets to add to a knowledge base from unstructured text, can be improved by including cross-lingual word embeddings [27], mainly due to the fact that knowledge bases of different languages can have different assertions, and embeddings offer a way of sharing that information across language boundaries.

The work on learning cross-lingual text representation has also resulted in a surprising finding, namely that embedding spaces have similar shapes across languages [28]. It is possible to train a machine translation model without sentence-aligned training data, and only using monolingual datasets of the respective languages to train embeddings, and the cross-lingual alignments between produced by an unsupervised approach of matching the two embedding distributions.

Despite embeddings being well suited to representing words either in a monolingual or a multilingual setting, most such configurations have a weakness in representing different word senses [29]. We can counter this by modeling them as distinct entries in the joint word-sense embedding space, improving their performance on the sense clustering problem. Therefore, defining the embedding space correctly lets us learn text representations better suited for named entity disambiguation problems.

To summarize, cross-lingual word embeddings have capitalized on the recent methodological and algorithmic advances in the field of representation learning, owing their wide use to their simplicity, efficiency and handling of large corpora, as well as their solid and robust performance on different tasks.

### 2.1.2 Named entity disambiguation

Problems analogous to entity disambiguation appear in multiple domains where structure and consistency of information is insufficiently enforced for various reasons, leading to duplication and ambiguity. The theoretical foundations of entity disambiguation are defined in the theory of record linkage [30]. Related challenges can also be found in database integration [31, 32], object identification [33], duplicate detection [34] and word sense disambiguation [35, 36].

While in general, entity disambiguation can involve both duplicate entries in a database, as well as their mentions in text that are inherently ambiguous, this topic focuses on the latter situation. When observing our problem statement from a natural language processing perspective, we can describe our approach as disambiguation using background knowledge, a pattern, often found in literature [37, 38, 39]. In general, background knowledge base is structured in the form of graph or relational data, being amenable to network algorithms. For example, a PageRank score over a candidate meaning network can be used to estimate the correct meanings [40]. It can also be posed as a Markov logic problem [41].

The schema of the knowledge base determines how we calculate relational similarity. When the schema supports different types of relationships that occur with different frequencies, that alone can be an important disambiguation signal. Literature in quantifying these diverse relationships suggests an adaptive method of determining relational significance [42].

When solving the entity disambiguation problem in the generalized sense, the usual approach involves performing graph clustering over the entity graph using a certain similarity criterion [43]. In context of relational data, it is a combination of attribute similarity and relational similarity.

When defining the task of entity disambiguation in the scope of cross-linguality, we are faced with the assumption that all the content is in the same language: both the document and the disambiguation models and the knowledge base. When operating across multiple languages, the simplest solution for cross-lingual entity disambiguation is the one that simply disregards the language mismatch and tries to use the full textual content to perform the context similarity without any additional processing [44]. The authors have shown that using a merged bilingual knowledge base performed better than using just the document language knowledge base, mainly due to better domain coverage, but it performed much worse than a monolingual scenario. Another simple baseline uses the equivalent of just using the context-independent 'mention popularity' measure, backed by a dictionary [45]. We can construct the dictionary from anchor texts from non-English to English Wikipedia pages. An ideal system would be the one that would simply translate the document in the desired language and perform the disambiguation on the translation. However, it is frequently the case that named entities themselves cause issues to such systems, unless specfically handled. While reproducing this scenario manually is not feasible for our task, one may use machine translation to do this [46]. While the authors achieve up to 94% performance of a monolingual baseline, machine translation greatly complicates and slows down the processing, opening a window of opportunity for more efficient approaches.

## 2.2 Language in Social Media

The analysis of language in social media is a particular domain of interest since it relays not just the informational content of the posts themselves, but also gives us an insight of behavior on social media. In this topic, we focus on two particular problems: summarizing social media responses to news, and robustness of sentiment analysis across multiple domains.

### 2.2.1 Behavior in social media

There is an increasing amount of literature on social media sampling and summarization, where the goal is respectively to select a representative sample of messages on a given topic [47, 48] or to build a summary of an event from social media posts [49, 50, 51].

The idea of interestingness is also related to the work done on predicting re-posts of messages , where *retweets* are considered a proxy for interestingness [52]. The authors used a varied set of features for predicting retweets, including emotion-related information, the presence of lexical indicators of interestingness and mood indicators; bag-of-words features were also explored, but revealed low predicting power due to sparsity issues. An important application of interestingness modeling is predicting the spreading of ideas [53], where a combination of content, temporal and topological features were most important in predicting whether a message will get re-posted.

While people consider re-posting the title of the news item informative, research [54] shows that recent text summarization techniques produce content that can increase user engagement, since they reflect a balance between informativeness and interestingness, even when the summarized content is different from the tweets delivered by the news providers.

Interestingness is also explored in the context of whole conversations within social media [55], where it is decomposed into two modalities: the interestingness of both the participants and the conversation. While we can characterize the former by participation in interesting conversations, the latter is characterized by the topic, the communication dynamics of the participants and the conversation itself. Another aspect that we consider as part of interestingness is content quality, one of the most important indicators in question-

answering communities [56], where the model combines both the social aspects and the linguistic quality of the messages.

Work in this chapter has been inspired by the outcomes of studies analyzing the different types of social media reactions to different types of events, e.g. by the observation that breaking news is often accompanied by informational messages, whereas other news items are characterized by more conversational reactions [57]. Furthermore, events can elicit different levels of excitement and participation, depending on both the nature of the event (life, work), the characteristics of the participants [58] and even their opinion [59]. These multiple facets have been effectively tackled by multi-objective optimization approaches [60]. Recently, significant attention has been directed at detecting and tracking events as they surface in microblogs [61, 62]. Social media messages have also been used to predict the popularity aspect of news stories [63], as well as assist in their summarization [64].

### 2.2.2 Sentiment analysis across domains and languages

Sentiment classification is an important part of our information gathering behavior, giving us the answer to what the people's attitude is toward about a particular topic. It is also one of the natural language processing tasks which is well suited for machine learning, since it is a three-class classification problem, classifying every example into either positive, neutral, or negative. Earlier work applied sentiment classification to online datasets, such as product, restaurant, and movie reviews [65], training a model for predicting whether a particular review rates a movie positively or negatively. While in the review domain all examples are inherently either positive or negative, other domains that have appeared later may also deal with non-subjective content that does not carry any sentiment, especially on media such as web pages, online news, discussion groups, and other social media outlets. Separating subjective from objective examples has proven to be an even more difficult problem than separating positive from negative examples [66].

Another difficult problem in this area is dealing with different topics and domains: models, trained on a particular domain do not always transfer well onto other domains. While the standard approach is to use one of widely used classification algorithms such as multinomial Naïve Bayes [65], explicit knowledge transfer approaches have been proven to improve performance in these scenarios, such as using sentiment lexicons [67] or modifying the learning algorithm to incorporate background knowledge [68]. Some challenges are also domain-specific. For instance, while social media content contains a lot of expressed sentiment, the language is informal, affecting the performance by increasing the sparsity of the feature space. On the other hand, the patterns arising in informal communication, such as misspellings and emoticons, can be themselves used as signals. It has also been shown that within social media, using different document sources, such as blogs, microblogs and reviews, can improve performance compared to using a single source [69].

# Chapter 3

# Leveraging Knowledge Bases for Cross-Linguality

This chapter discusses the role of user-generated knowledge bases for coming up with a text representation that is robust enough to build cross-lingual models. We discuss the role of knowledge bases in cross-linguality. A *knowledge base* is a collection of entities, where every entity has its identifier, a set of labels, and a set of descriptions in multiple languages. We present several problems involving knowledge bases in an application to natural language processing.

## 3.1 Introduction

Having multi-lingual knowledge bases assumes that we can express the same concept in different languages. However, getting the correct meaning of a sentence in the first place is a hard problem even in a monolingual setting, without introducing the complication of translating it to another language. Therefore, we use cross-linguality as a strategy of finding shortcuts to perform natural language processing tasks over multiple language without the effort of understanding or fully translating the input.

For example, if we are interested in the topic of the document, we more or less only need to observe its word distribution to achieve adequate performance. Also, if we are interested in what named entities are being mentioned in a document written in another language, we only need to perform named entity extraction and disambiguation across them, provided that the knowledge base in question is capable of inferring that mentions in different languages correspond to the same entity. Knowledge bases, such as Wikipedia or JRC-Names [70] satisfy that requirement. Presumably, neither of these examples should require machine translation for the system to perform well.

### 3.1.1 Use cases

First, we try to use multilingual descriptions of entities in the knowledge base as a weakly-aligned comparable corpus and use it as a translation model despite having sparse and coarse alignments, using the role of the knowledge base as training data for cross-linugual dictionary induction and similar use cases.

Second, we try to use the entities themselves as a representation domain, where we can represent every document as a set of the entities it mentions, where the output is a connection between a document and its entities. We refer to the process of obtaining these connections as entity resolution. Here, the role of the knowledge base is to represent the document so that it is easier to manage and display due to more explicit semantics.

In both cases, the text annotation is an exercise in identifying the semantics of the content. For instance, when dealing with a corpus of plain text, the basic interactions in this medium are reading, browsing, and search. If we introduce cross-lingual connections among related documents that may be in different languages, we can see possible translations and related content, thus opening the possibility of using another axis for interaction. Furthermore, introducing named entities gives another dimension to the medium: we have another meaningful way to browse and aggregate the content, especially when trying to piece together temporally disconnected stories [71]. The same can be said for introducing topics, recommending related content, and adding social interactions. After all, the premise of social media services is that the viewers of such services do not only read and consume the content, they also create and interact with it.

### 3.1.2 Multi-lingual language resources

While sentence-aligned corpora are extensively used in machine translation, comparable corpora have attracted attention due to the difficulties (hard criterium of sentence-alignment) in collecting parallel corpora, especially for under - resourced languages and domains. In recent years, the use of cross-lingual comparable corpora has attracted considerable attention in the natural language processing community [72].

A common denominator among many natural language processing tasks, the bag-of-words model is commonly used as a baseline feature representation. In many tasks involving multi-linguality, we are faced with the constraint that meaning and knowledge cannot be transferred easily when we consider words as independent units. The concept of cross-linguality assumes that we can express the same meaning in different languages. However, actually getting the correct meaning of a sentence is a hard problem even in a monolingual setting, without introducing the complication of translating it to another language.

Imagine we had the ability to translate a document from language $a$ in a bag of words representation into a bag of words in another language $b$. This way, we can re-use a natural language processing (NLP) component for language $b$ to process examples from class $a$, provided that the said NLP component takes its input data as bag-of-words.

This allows us to relax the requirements for the language resources: we are using comparable corpora, which can be easier to obtain and are more diverse than parallel corpora. Specifically, it allows us to consider Wikipedia as such a resource [21], since parallel corpora are relatively scarce, especially for technical and niche domains, and for language pairs not involving English.

The properties that we attempt to exploit are the following:

- **Availability of weakly aligned data.** While machine translation literature employs monolingual language models in the decoding phase, we use the local word co-occurrence patterns across large corpora as means to contextualize the data, while properly aligned bilingual corpora are rare, there are still plenty of weakly aligned resources, such as Wikipedia, Wikidata, and the Linked Open Data ecosystem.

- **Non-linearities in cross-lingual relationships.** While a dictionary mapping, a simple example of a linear translation, can provide a reasonable baseline, it often fails to capture contextual nuances that translate poorly across languages. We hypothesize that these nuances do not translate well when using linear mappings, since word meanings do not always have one-to-one correspondences across languages.

- **Sparsity of dictionary induction.** From the perspective of their linear algebra representation, translation models discussed in this chapter produce dense outputs. However, when the input document has a small amount of words, we can reasonably

expect the same from the output document — it rarely happens that the translation is more verbose than its input. For this purpose, we leverage the work on compressive sensing and re-imagine translation as a lossy communication channel from which we are trying to reconstruct the phrase in the target language.

### 3.1.3 Dataset description

We train our models on various language versions of Wikipedia. Since many documents in Wikipedia have cross-lingual links that determine correspondence of the same article in other languages, we use those cross-lingual links as alignments of individual data points between two languages.

Table 3.1: Number of aligned documents among selected languages of Wikipedia.

| Source language | EN | ES | DE | SL | CA | HR |
|---|---|---|---|---|---|---|
| Document count | 4212479 | 968603 | 1523542 | 136313 | 395185 | 121511 |
| Links with EN | | 611093 | 767575 | 80326 | 265757 | 81897 |
| Links with ES | | | 322478 | 39932 | 250127 | 54409 |
| Links with DE | | | | 51699 | 159447 | 60078 |
| Links with SL | | | | | 33370 | 31328 |
| Links with CA | | | | | | 38333 |

We use the English, Spanish, German, Catalan, Slovene and Croatian versions of Wikipedia. The set of languages was motivated by the research project that this work was part of. The alignment sizes among them are presented in Table 3.1.

### 3.1.4 Summary of scientific contributions

1. First, Section 3.2 characterizes and evaluate a set of approaches for scalable dictionary induction for textual content using comparable corpora as training sets. The focus is on finding ways to use monolingual source corpora in tasks involving multilingual content, such as dictionary induction and cross-lingual mapping. We pose the dictionary induction problem as a problem of finding a projection operation from two comparable corpora into a common cross-lingual space that is used as an intermediate representation when translating from one language to another.

2. We propose a scalable method for cross-lingual dictionary induction base using distributional representations of words as a representation layer and kernel approximation for the non-linear translation layer. We show that the non-linear mappings obtained by kernel approximation can outperform linear mappings on the task of bilingual dictionary induction while still maintaining sufficient scalability properties. We also characterize the performance given different model complexity levels. The cross-lingual similarity estimates produced by the model correlate well with human judgments.

3. In Section 3.3, we present another side of the problem of translating vector space text representations from one language to another using only comparable bilingual corpora, focusing on the sparsity properties of translation. The approach uses Elastic Net regularization [73] on a multivariate regression using low-rank Canonical Correlation Analysis [74] to achieve sparsity in favor of a separate post-processing thresholding step. Results show the importance of result sparsity both in cross-lingual information retrieval and fit of dictionary induction.

4. In Section 3.4, we present another use case of using knowledge bases to better understand text, this time focusing on representing it as a collection of named entities mentioned in thext. We formulate this as an entity resolution problem, where we are trying to choose the correct corresponding entities from the ontology for the entities mentioned in text. We propose a solution to that problem that uses the diverse relationships between entities in the knowledge base as the key disambiguation signal. We study various methods for estimating relatedness between entities, used in collective entity resolution. We define a unified entity resolution approach, capable of using implicit as well as explicit relatedness for collectively identifying in-text entities. As a relatedness measure, we propose a method that expresses relatedness using the heterogeneous relations of a domain ontology. We also experiment with other relatedness measures, such as using statistical learning of co-occurrences of two entities or using content similarity between them. Evaluation on real data shows that the new methods for relatedness estimation give good results, both in monolingual and cross-lingual settings.

## 3.2 Cross-Lingual Document Similarity Estimation and Dictionary Induction with Comparable Corpora

This section describes an approach for performing bilingual dictionary induction even when trained on widely available comparable bilingual corpora. We also show its capability to provide cross-lingual similarity estimates that correlate well with human judgments. We implement an approach using a nonlinear bilingual translation model that we train using comparable corpora. We propose a method using word embeddings and kernel approximation to train scalable non-linear transformations. We demonstrate that this novel method works better on a majority of evaluated language pairs. This chapter is an extended version of the work published in [75].

### 3.2.1 Introduction

The bag-of-words document representation allows us to somewhat relax the requirements for the language resources: we are using comparable corpora, which can be easier to obtain and more diverse than parallel corpora. Specifically, it allows us to consider weakly-aligned multilingual corpora, such as Wikipedia. This additional relaxation on the input structure constraints allows us greater flexibility in translation or dictionary creation tasks, especially when bootstrapping under-resourced languages. To accommodate this, we find a good representation of language that can efficiently map from one language space to another.

Besides learning a monolingual representation, cross-linguality also includes learning a cross-lingual mapping. Here, there are several possibilities on how to pose the problem: in the trivial case, if two languages are very similar, a simple dictionary mapping already goes far. However, with language pairs which have few similarities, the mapping may be more complex. This section focuses on the latter example, showing that upon learning the right representation, we can use linear models to perform the actual mapping.

The main scientific contributions of this section are the following:

- A novel method using kernel approximation on word embeddings that outperforms several related approaches on dictionary induction with comparable corpora on several language pairs.
- We demonstrate that this novel method works better on a majority of tested language pairs.
- A survey on several cross-lingual dictionary induction methods.
- We show that methods based on word embeddings can successfully be used to estimate cross-lingual document similarity.

Throughout the section, we use the following notation: $X$ and $Y$ represent the vector space representations of source and target language corpora, which are assumed to be aligned at the document level. Conversely, $x$ and $y$ are used to represent vector representations of individual documents from source and target languages, respectively. The problem is posed as finding a mapping $X \rightarrow Y$ that minimizes a certain error metric.

Since the goal of this paper is to provide a recipe for learning an optimal representation and mapping strategy, we evaluate the performance of the methods on the task of dictionary induction on several language pairs.

Scenarios for bilingual dictionary induction have traditionally relied on machine translation (MT). The typical way of re-using NLP components across languages consists of translating the input content using an MT system, and using the translations as input to

the NLP component [4]. This pattern can be applied to problems such as cross-lingual entity linking [5] and sentiment analysis [6], as well as entity tracking, topic detection and cross-lingual recommendation [76].

However, using a full machine translation pipeline can be impractical, since the existence of such a system assumes that there are existing sentence-level alignments that were used to train such a system. However, many languages and vertical domains may not have sufficient aligned corpora for training machine translation systems, but may have access to comparable corpora. Therefore, we explore the consequences and possibilities of using comparable corpora for translation tasks.
From the performance viewpoint, machine translation represents an upper bound that we are unlikely to outperform by simpler approximations on a general domain.

Furthermore, as discussed in Ruder [15], Neural Machine Translation approaches implicitly learn a shared cross-lingual embedding space by optimizing for the Machine Translation (MT) objective, whereas this thesis focuses on models that explicitly learn cross-lingual word representations also for other use cases. These methods generally do so at a lower cost than MT and, in terms of speed and efficiency, can be considered to be to MT what word embedding models are to language modeling.

#### 3.2.1.1 Word embeddings

Recently, natural language processing tasks started relying on using word embeddings as the underlying representation for solving their problem, such as multilingual classification using word embeddings [77] without requiring word-level alignments.

The approach [8] poses the problem of translation as a combination of first learning a low-dimensional representation of the aligned corpora, and then calculating the translation matrix that translates one low-dimensional representation of text to another using a least-squares optimization problem, using a skip-gram model that learns a linear mapping from words to components using neural network learning techniques, maximizing the probability of a word given its neighborhood [20, 78]. Their experiments show that learning a linear mapping by solving a least-squares problem using stochastic gradient descent yields good results for predicting word translations. The skip-gram model with negative sampling approximates a language model but focuses on learning efficient word representations, rather than accurately modeling word probabilities.

The skip-gram model for a given language is defined in the following fashion: let $w_1, w_2, \ldots, w_T$ represent a sequence of training words. Its objective is to maximize the average log-probability of words' contexts within the training corpus of size $T$ given a training window of size $k$:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{k}\log p(w_{t+j}|w_t) \tag{3.1}$$

The same idea can also be applied the other way: the objective of the Continuous bag-of-words model (CBOW) is to maximize the log-probability of each word given its context:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{k}\log p(w_t|w_{t+j}) \tag{3.2}$$

In both cases, the probability model uses two parameter vectors for every word: $u_w$ and $v_w$ that represent the input and output vectors that map from the input to the hidden layer, and from the hidden to the output layer. Expanding to the entire vocabulary, they

can be seen also as matrices $U$ and $V$. The probability, given a vocabulary of size $V$ is defined as:

$$p(w_i|w_j) = \frac{\exp(u_{w_i}^T v_{w_j})}{\sum_{l=1}^{V} \exp(u_l^T v_{w_j})} \tag{3.3}$$

These models are trained using stochastic gradient descent, where the gradient is obtained from the propagation rule. The approach also uses a negative sampling technique to generate negative examples by randomly generating $k$ word-context pairs for every real word-context pair $w, c$ using a stochastic gradient descent approach.

The Skip-gram model learns word embeddings similarly to neural language models [79], but without a non-linear hidden layer. Literature [80] also shows that this approach is equivalent to factorizing a shifted positive point-wise mutual information matrix (SPPMI) into $U$ and $V$.

$$V^T U = \text{SPPMI}_k(w, c) = \max(\text{PMI}(w, c) - \log k, 0) \tag{3.4}$$

where PMI is defined as:

$$\text{PMI}(w, c) = \log \frac{p(w, c)}{p(w)p(c)} \tag{3.5}$$

Since the SKIP model outperforms the CBOW model in related tasks, we use the SKIP model in our experiments.

For our purpose, we are interested in the low-dimensional embedding $V$, which we can use to multiply with the input vector to obtain the representation in the embeddings space. Subsequently, we denote the word-embedding representation of the query document $x$ as $x_{low} = xV^T$.

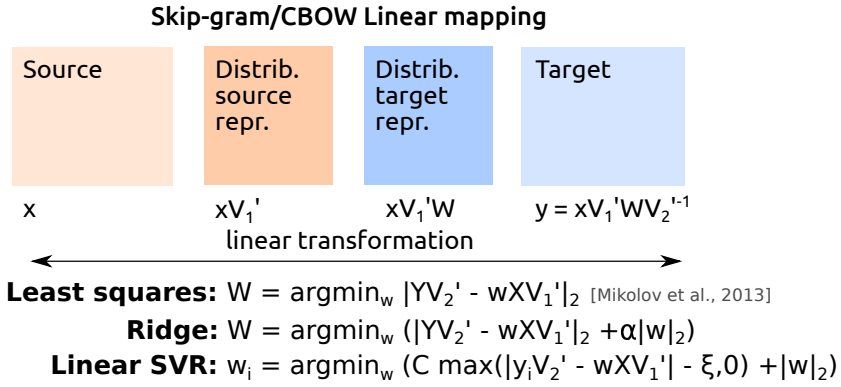


Figure 3.1: Integration of word embedding representations and various cross-lingual linear mappings. $V_1$ and $V_2$ denote the embedding parameters, $W$ denotes the translation matrix among the two embedding spaces.

Let $X_{low}$ and $Y_{low}$ denote low-dimensional word embeddings of two language corpora, obtained as hidden layer representations from either Skip-gram or CBOW models. In the approach in Mikolov, Le, and Sutskever [8], the authors propose to learn a linear mapping $W$ that gives the least squares solution by minimizing $|Y_{low} - WX_{low}|$. The approach uses a stochastic gradient solver to obtain a solution $W$, and is denoted with the LSQ-* prefix in our experiments. Figure 3.1 shows the architecture of this approach.

We can demonstrate this particular transformation by taking a word in the input space, transforming it to the embedding, and mapping it back to the word space. For instance, using the SKIP representation, the word *king* returns as *king, iii, queen, reign, throne* for its top ranking components, all of which are strong indicators of local context. When using representations using word embeddings in other approaches, we denote that using LSQ-SKIP for the skip-grams representation with least-squares mapping to translate between embeddings of different languages.

### 3.2.2 Proposed approach

This subsection presents several ways to improve on LSQ-SKIP. Given the described related work on word embeddings, we propose a novel method for performing fast non-linear approximations for mapping between language vector spaces. We evaluate it on several architectures with different representation learning models, as well as translation models.

While linear cross-lingual mapping models such as LSQ-SKIP can be efficiently estimated, that inherently limits the expressiveness of the model. We hypothesize that we can perform better transformation by first using representation learning with word embeddings on individual languages, but using a non-linear mapping from one language to another instead of a least-squares model.

However, we are still left with the constraint of being efficient on large corpora, scaling to hundreds of millions of tokens and millions of distinct words. Learning a non-linear mapping between two large vector spaces is computationally difficult, which is why we resort to approximations. To that end, we hypothesize that due to the embedding space being robust to small localized noise, we can use a faster, approximate cross-lingual mapping model.

#### 3.2.2.1 Monolingual embeddings with a kernel approximation mapping

In order to maintain a usable run time, most approaches mentioned so far rely on providing a linear projection for translation, potentially limiting performance. We introduce non-linearities among the embedding spaces either by using different kernels in the support vector regression, but the number of data points and output dimensions can often be too big for a quadratic training algorithm. In this subsection, we propose solving the task of mapping between word embeddings using an approach that can efficiently represent non-linearities. We use the Nyström method for kernel approximation [81, 82] that speeds up learning with very large datasets by using an approximate kernel map. Instead of training a non-linear regression, we can apply an approximate kernel map to our input and use much more efficient linear solvers downstream to produce the outputs, such as stochastic gradient descent [83].

The Nyström kernel approximation approximates the full kernel matrix $K$ by first sampling $m$ examples, denoted by $\hat{x}_1, \dots \hat{x}_m$, and then constructs a low rank matrix $\hat{K}_r = K_b K^+ K_b^T$, where $K_b = [\kappa(x_i, \hat{x})]_{N \times m}$, $\hat{K} = [\kappa(\hat{x}_i, \hat{x}_j)]_{m \times m}$, and $K^+$ is the Moore-Penrose pseudo-inverse of $\hat{K}$ and $r$ is the rank of $\hat{K}$. Having the approximation $\hat{K}$ and a kernel function $\kappa$, we define the transformation

$$z(x) = \hat{\Sigma}_r^{1/2} \hat{U}_r (\kappa(x, \hat{x}_1), \dots, \kappa(x, \hat{x}_m))^T \tag{3.6}$$

that becomes an $m$-dimensional representation of the input $x$. Learning a linear machine that predicts $Y$ using $z(X)$ is equivalent to using the kernel $\kappa$ directly with a SVM regression that predicts $Y$ directly from $X$.

The approximation is achieved by sub-sampling the data on which the kernel is evaluated, allowing us to represent new samples by the sigmoid kernel values to those samples,

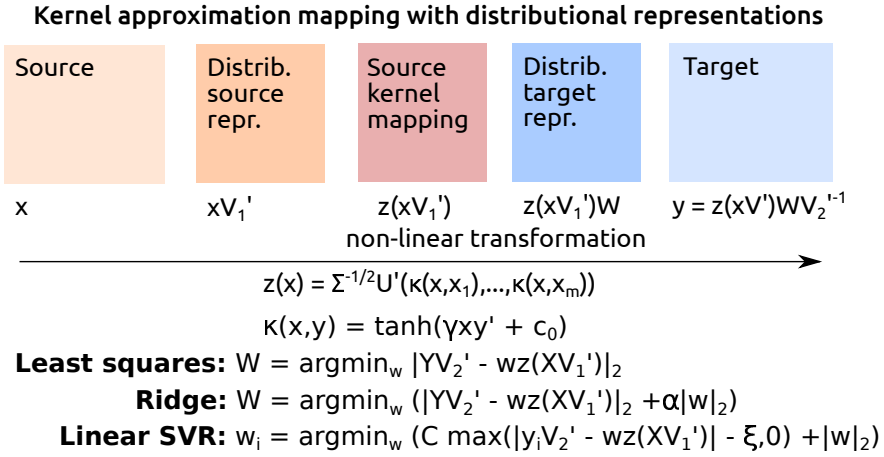


Figure 3.2: Integration of kernel approximation as a non-linearity in the cross-lingual translation pipeline. Note that this architecture only supports one direction. $z(x)$ denotes the non-linear prediction operator, while $V$ denote the embedding parameters and $W$ the translation matrix from kernel space to the target embedding space. $C$ is a regularization parameter for the SVM model.

and running a fast SGD-based multivariate ridge regression on the output [84]. Since the running time for kernel approximation is proportional to the number of samples $m$ and number of dimensions $d$, this allows for linear scalability with respect to training dataset size.

Figure 3.2 and Algorithm 3.1 shows the training approach: first, word vectors are translated into embedding vector in both languages. Second, using the word embeddings corpus of the source language $X_{low} = XV_a^{\intercal}$, we generate a kernel approximation model around random documents, denoted as $X_{ka} = z(X_{low}; \Theta)$. Finally, we fit a multivariate least squares regression from the first language kernel space to the target language embedding, solving: $|Y_{low} - WX_{ka}|_2$.

At prediction time, as shown in Algorithm 3.2, given $V_1$, the non-linear function $z$ with parameters $\Theta$, and $V_2$, we can then produce translations of individual documents vectors, as shown in the equation in Figure 3.2.

#### 3.2.2.2 Monolingual embeddings with an ExtraTrees mapping

In order to assess the compromise of kernel approximation, we also consider other regression methods to achieve the same goals. For instance, using a multivariate regression, composed of Extremely randomized trees (Extra-Trees) [85], which we denote by EXT-*. While this method has been proposed for being fast to train compared to other decision tree learning approaches, it still has a higher run-time compared to kernel-approximation based methods, we hypothesize that it can also generate improvement over the linear methods. In essence, we train a separate regression tree for every embedding dimension of the target language, given the embeddings of the source language.

#### 3.2.2.3 Monolingual embeddings with a cross-lingual CCA mapping

A family of cross-lingual information retrieval also poses this problem as a CCA (canonical correlation analysis) optimization that gives a common space between the embedding-based representations of two languages. This approach does not try to directly map from one

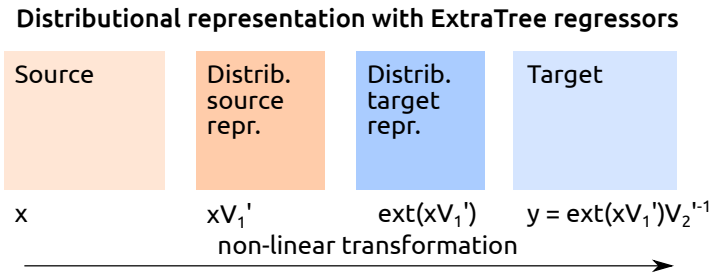


Figure 3.3: Multivariate ExtraTree estimators as a non-linear cross-lingual mapping. The *ext* operator represents the multivariate ExtraTree model obtained by concatenating predictors for individual dimensions in the target embeddings space. $V_1$ and $V_2$ correspond to the embedding parameters.

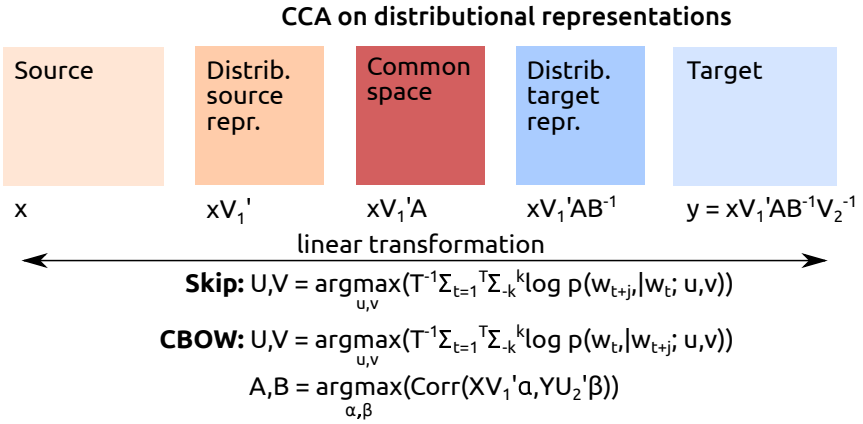


Figure 3.4: Monolingual embeddings with a CCA cross-lingual transform (CCAT-SKIP and CCAT-CBOW). $V$ and $U$ denote the embedding parameters, and $A$ and $B$ denote the CCA mappings that map from both languages to the common space.

**ALGORITHM 3.1:** Training of monolingual embeddings with a kernel approximation mapping model

**Input**: source language training corpus $X$, target language corpus $Y$
**Input**: alignment indicator matrices $A_1$ and $A_2$
**Output**: source and target embedding models $V_1$, $V_2$
**Output**: kernel approximation parameters for target embedding dimension $d$: $\Theta_d = (\Sigma_d, U_d, \hat{X}_d)$
**Output**: kernel space to target embedding space mapping $W$

```
// Train embedding representations
```
$W_X \leftarrow$ train skip-gram$(X)$
$W_Y \leftarrow$ train skip-gram$(Y)$
```
// Select aligned data points from both sides
```
$X_{aligned} \leftarrow XA_1$
$Y_{aligned} \leftarrow YA_2$
```
// Transform to low-dimensional embedding space
```
$X_{low} \leftarrow V_1 X_{aligned}$
$Y_{low} \leftarrow V_2 X_{aligned}$
**for** *dimension* $d$ *of* $Y_{low}$ **do**
    $\hat{X}_d \leftarrow \text{choice}(X_{low}; m)$ `// Sample m datapoints`
    $\hat{K_d} \leftarrow [\kappa(\hat{x}_{di}, \hat{x}_{dj})]_{m \times m}$ `// Calculate a kernel matrix from them`
    $U_d, \Sigma_d, V_d \leftarrow \text{SVD}(\hat{K_d})$ `// Decompose the kernel matrix`
    `// Transform from source embedding space to kernel space`
    $x_{ka,d} \leftarrow \Sigma_d^{1/2} U_d [\kappa(x_{low}, \hat{x}_{d,0}), \ldots, \kappa(x_{low}, \hat{x}_{d,m}]$
**end**
```
// Train model projecting from kernel space to target embedding space
```
$W \leftarrow \arg\min_W |Y_{low} - WX_{ka}|_2 + |W|_2$

representation to another, but instead attempts to learn a common space using CCA. The resulting common representation allows us to map across langauges by means of linear operations. Besides providing a mapping, it also allows for performing other operations in this common space, such as cross-lingual document retrieval or cross-lingual document linking. In other words, it is a combination of the cross-lingual mapping approach in LRCCA with embeddings instead of bag-of-words. While LRCCA uses the cross-covariance matrix decomposition to obtain the low-rank transforms, CCAT-* uses the Skip-gram representation for each langugage, denoted CCAT-SKIP, as illustrated in Figure 3.4.

### 3.2.3 Baselines and related approaches

The main hypothesis of this work is that using an unsupervised representation learning step can improve vector space translation at a given translation model complexity. We explore performance of several supervised and unsupervised representation learning approaches on aligned document pairs from two natural languages.

#### 3.2.3.1 Cross-lingual Latent Semantic Indexing

One of the early approaches that deal with obtaining a low-dimensional text representation for use in multilingual contexts is cross-lingual Latent Semantic Indexing (denoted as CLLSI) [86], a method that has been known to work well for cross-lingual information retrieval [19], leveraging the idea of latent concepts. It is a supervised approach that performs latent semantic indexing on aligned document pairs by concatenating the feature spaces of both languages, weighted using the TF-IDF scheme. It simultaneously works as a translation model as well as a representation model.

CLLSI solves the polysemy and synonymy problems by applying SVD to compute the

**ALGORITHM 3.2:** Prediction with monolingual embeddings with a kernel approximation mapping model

**Input**: example in source language $x \in X$
**Input**: source and target embedding models $V_1$, $V_2$
**Input**: kernel approximation parameter model for target embedding dimension $d$
$\Theta_d = (\Sigma_d, U_d, \hat{X}_d)$
**Input**: kernel space to target embedding space mapping $W$
**Output**: prediction in target language vector space $y_{pred}$
$x_{low} \leftarrow V_1 x$ `// Transform to source embedding space`
**for** *dimension d of* $y_{low}$ **do**
  `// Transform from embedding to kernel space`
  $x_{ka,d} \leftarrow z(x_{low}; \Theta_d) = \Sigma_d^{1/2} U_d[\kappa(x_{low}, \hat{x}_{d,0}), \ldots, \kappa(x_{low}, \hat{x}_{d,m}]$
**end**
$y_{low} \leftarrow W x_{ka}$ `// Transform to target embedding space`
$y_{pred} \leftarrow V_2^{+} \hat{y}_{low}$ `// Transform to target word space`

approximation matrix to the term-document matrix by decomposing the approximation matrix into three matrices, as shown in Figure 3.5. In this process, the high-dimensional data set is reduced to a lower dimensional vector space, preserving the substructure of the original data while reducing the amount of variations.

The approach involves concatenating the $X$ and $Y$ matrices by its alignment so that every example contains features from both languages in the same row, resulting in a matrix $Z$:

$$Z = [XY] \tag{3.7}$$

Taking Wikipedia versions of two languages as an example, the columns corresponding to $X$ would be bag-of-words representation of the document in the first language pair, and $Y$ its counterpart in the other language.

We define the bijective operators $M(x) : X \to Z$, and $N(y) : Y \to Z$ which simply shift the dimensionality of the inputs so that all dimensions of $X$ come before the dimensions of $Y$. In matrix form, $M$ and $N$ can be implemented as concatenations of identity matrices with empty matrices.

Then, we perform latent semantic indexing to obtain a low-rank representation of the covariance matrix of $Z$, with $U$, $\Sigma$ and $V$ being the factorization results of the singular value decomposition of $Z^\intercal Z$:

$$Z^\intercal Z \approx U \Sigma V^\intercal \tag{3.8}$$

This gives us the ability to represent every example in a low-dimensional space, so that $\hat{Z} = ZU$

The translation can then be performed by applying a combination of singular value decomposition factors and the $M$ and $N$ language masks:

$$\Psi_{lsi}(x) = N^{-1} M x V^\intercal U \tag{3.9}$$

#### 3.2.3.2 Low-rank Canonical Correlation Analysis

Another family of methods that is frequently encountered in cross-lingual text analysis scenarios is Canonical correlation analysis (CCA) that addresses the following optimization problem: given two comparable corpora matrices $X$ and $Y$, find a pair of linear mappings that maximize the correlation between variables of $X$ and $Y$. Those mappings can be

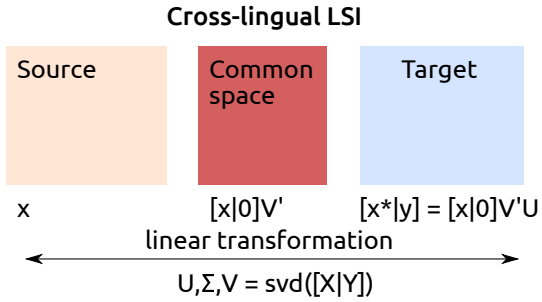


Figure 3.5: Architecture of Cross-lingual LSI, showing how the target $y$ is obtained from the input $x$

used to translate examples from both spaces into a common subspace. The intuition behind Low-rank CCA Translation (LR-CCA) [25] is that the cross-lingual mappings can be more compactly represented in a low-dimensional basis instead of the high-dimensional word spaces, and that simply concatenating the document pairs, as done in approaches, such as CLLSI, can lead to loss of information. LRCCA is a combination of two steps: first, a low-rank SVD decomposition of the co-variance matrix of both aligned corpora is performed, which acts as the representation learning step. Given these transforms, we perform canonical correlation analysis on the low-rank representations of both languages to learn a translation model between them. The input data uses the TF-IDF weighing scheme.

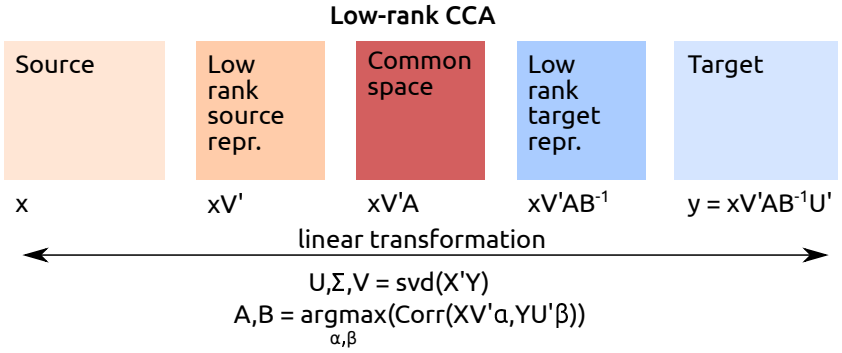


Figure 3.6: Architecture of the low rank CCA bidirectional translation pipeline.

Figure 3.6 illustrates the various representation spaces a document $x$ can be transformed to. We denote the transformation from the source language word space to the low-rank space with $V$ and $U$ for the source and target languages respectively, and $A$ and $B$ to represent the mapping from both low-rank spaces to the shared representation space. While this method has a stronger model for cross-lingual correlations than CLLSI, it is still limited for tackling polysemy.

The method first decomposes the cross-covariance matrix using singular value decomposition (SVD):

$$X^\intercal Y \approx U \Sigma V^\intercal \tag{3.10}$$

This step enables us to first map the examples from the $X$ and $Y$ spaces to a pair

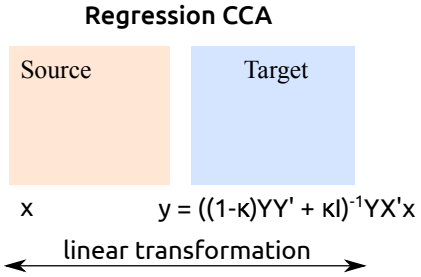


Figure 3.7: Illustration of Regression CCA, showing the direct linear transformation expression. $\kappa$ denotes the regularization strength.

of lower-dimensional spaces $\hat{X} = XV^\intercal$ and $\hat{Y} = YU$. Next, we apply standard CCA on $\hat{X}, \hat{Y}$, obtaining the linear mappings $A$ and $B$ that map from individual input spaces in the common space, so that $\hat{X}A = \hat{Y}B$.

Figure 3.6 illustrates the various representation spaces a document $x$ can be transformed to, from the source representation in TF-IDF-weighted bag-of-words, towards a low-rank representation $xV^T$, and a common space representation $xV^TA$. Given that $xV^TA$ should be maximally correlated with $yU^TB$, where $y$ is the corresponding vector in the target language, we can approximate $y$ by applying the entire chain of linear transformations. Since $B$ may not be invertible and the entire chain needs to be computed beforehand, we employ the Moore-Penrose pseudo-inverse [87] $B^+$ to approximate $B^{-1}$ instead of running a linear system solver on a per-query basis. Finally, we compute the final projection as $y = xV^TAB^+U^T$.

#### 3.2.3.3 Regression CCA

We also consider other text representations using CCA-based approaches for translation, such as Regression Canonical Correlation Analysis (CCAR). Regression CCA [24] is a method that specializes in computing an accurate mapping for a single input example $x_i$, showing good results on the task of cross-lingual information retrieval, where the goal is to predict which document in the other language is the correct alignment.

However, since it uses a least-squares approach for obtaining the solution, the output is still a dense vector. The approach is formulated as follows: let $X$ and $Y$ be aligned corpora, represented in TF-IDF vector spaces, $Q$ be the covariance matrix of $Y$, regularized by the hyper-parameter $\kappa$ controlling the regularization strength:

$$Q = (1 - \kappa)Y^TY + \kappa I \tag{3.11}$$

Given an input $x$, we solve the following optimization using a conjugate gradient solver to obtain $g(x)$:

$$g(x) = \arg\min_{g^*} |Qg^* - YX^Tx| \tag{3.12}$$

Then, we normalize the obtained $g(x)$ to obtain the final solution $\hat{y}$:

$$\hat{y} = -\frac{g(x)}{\sqrt{q^TXY^Tg(x)}} \tag{3.13}$$

Since the method works on a per-query basis, it works well also on infrequent words, due to the fact that it iteratively optimizes a solution around a given word. On the other

hand, the per-query conjugate gradient solver is computationally very demanding. When mapping the same word across and back, the reconstruction contains terms that are more indicative of the topic. For example, *king* reconstructs as *king, kingdom, royal, reign, son, crown*. While the topical indicators can be beneficial for tasks, such as information retrieval, we hypothesize that it may be better to model a finer grained context instead.

### 3.2.4 Experiments

To summarize, we want to test the following hypotheses:

- Do approaches using word embeddings outperform baselines?
- Do non-linear approaches outperform linear approaches?
- Can we can get the same level of performance when using approximate non-linearities approaches compared to exact non-linear estimators?

We train our models on various language versions of Wikipedia. Since many documents in Wikipedia have cross-lingual links that determine correspondence of the same article in other languages, we use those cross-lingual links as alignments of individual data points between two languages. In turn, this defines that the alignment is on the document level, which assumes that the pair of articles has corresponding content. The datasets used for this evaluation are the English, Spanish, German, Catalan, Slovene and Croatian versions of Wikipedia. We enumerate the approaches in Table 3.2. The ordering and segmentation of the approaches corresponds to the hypotheses and experiments we wish to perform.

Table 3.2: An overview of all methods used in this evaluation, showing the differences in representation learning and cross-lingual mapping components, where applicable. The methods' segments represent the hypothesis that they are designed to test.

| Method name | Representation learning | Cross-lingual mapping |
|---|---|---|
| **Baselines** | | |
| CLLSI [86] | LSI on concatenated language matrices | |
| LRCCA[25] | SVD on cross-cov. matrix | CCA |
| CCAR [24] | CCA via conjugate gradient solver | |
| **Word embeddings, linear mapping** | | |
| LSQ-SKIP [8] | Skip-gram | Least squares |
| CCAT | Skip-gram | CCA |
| **Word embeddings, approximate non-linear mapping** (proposed approach) | | |
| SIGLSQ-SKIP | Skip-gram | Kernel approximation, least squares |
| **Word embeddings, non-approximate non-linear** | | |
| EXT-SKIP | Skip-gram | Extremely randomized tree regression |

#### 3.2.4.1 Evaluation dataset

Since the different language versions of Wikipedia have differing sizes and not strictly comparable documents, the training set for the cross-lingual mapping consists of only the documents present in both languages used in the comparison. For steps involving monolingual representation learning, the we use the entire dataset.

When pre-processing the data for learning, we cut off low-frequent words, so that the remaining words represent 95% of the weight of the matrix. We tuned the hyper-parameters

of the methods, such as regularization and kernel coefficients using cross-validation, except for the number of components, where the experiments were performed for different numbers of components, ranging from 100 to 600 with an increment of 50. In the results, we report the result of the component dimensionality that we chose based on the highest performance on a separate validation set.

To evaluate the performance we use the DBPedia version of Wiktionary[1] in its RDF form [88] that has been further linked via crowdsourcing. The purpose of using dual benchmarks is that they offer coverage on slightly different language pairs.

We use these dictionaries as a benchmark for our vector space translation models. After training the translation models on a comparable bilingual corpus, we ask it to predict a translation of a given word, feeding it a sparse input with a weight of 1.0 in the dimension of the input word. The prediction of the model is a feature vector in the target language with a corresponding weight vectors. Assuming that the correct translation would have the highest weight in the output, we sort the output dimensions by value. We also measure precision-at-10, the observed likelihood that the correct word was in the top ten dimensions ordered by the predicted weight. Alternatively, we measured the mean reciprocal rank of the correct word, which is denoted by MRR. We omit those results for brevity, since they lead to the same conclusions as results measuring precision-at-10.

The evaluation explores the space of various cross-lingual mapping mechanisms, representation learning mechanisms and their parameter spaces for the task of dictionary induction.

#### 3.2.4.2 Experiments on bilingual dictionary induction

For evaluation scenarios using Wiktionary, we use only the language pairs that had more than 500 shared terms in their bilingual dictionary. The test set consists of 2000 words for every language pair. While some language pairs have more word pairs, we restricted the size to keep the evaluation dataset sizes comparable among language pairs.

When pre-processing the data for learning, we cut off infrequent words, so that the remaining words represent 95% of the weight of the matrix in order to make the training and pre-processing less computationally intensive. Varying this ratio did not significantly change the behavior of the tested models. Hyper-parameters of the methods were tuned using cross-validation on a tuning set of 200 Wiktionary word pairs.

We use the dictionaries as benchmark for our vector space machine translation models. Every dictionary entry pair is translated into a pair of document vectors. After training the translation models on a comparable bilingual corpus, we ask it to predict a translation of a given word. To evaluate, we measure precision at top-10.

Table 3.3: Precision-at-10 of different method on different language pairs on dictionary induction.

| Method | es-ca | es-de | es-en | ca-es | ca-de | de-es | de-ca | de-en | en-es | en-ca | en-de | en-hr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CCAR | **0.89** | 0.03 | 0.08 | **0.89** | 0.05 | 0.06 | 0.03 | 0.02 | 0.04 | 0.02 | 0.01 | 0.10 |
| CLLSI | 0.26 | 0.13 | 0.11 | 0.15 | 0.05 | 0.08 | 0.04 | 0.08 | 0.14 | 0.15 | 0.04 | 0.06 |
| LRCCA | 0.48 | **0.31** | 0.18 | 0.35 | 0.11 | 0.25 | **0.10** | **0.20** | 0.47 | **0.32** | **0.10** | **0.20** |
| LSQ-SKIP | 0.62 | 0.18 | 0.18 | 0.50 | **0.12** | 0.29 | 0.09 | 0.10 | 0.52 | 0.21 | 0.08 | 0.08 |
| CCAT-SKIP | 0.29 | 0.10 | 0.11 | 0.42 | 0.06 | 0.15 | 0.03 | 0.03 | 0.25 | 0.04 | 0.02 | 0.04 |
| SIGLSQ-SKIP | 0.69 | 0.30 | **0.20** | 0.50 | 0.10 | **0.30** | **0.10** | **0.20** | **0.60** | 0.20 | **0.10** | **0.20** |
| EXT-SKIP | 0.69 | 0.30 | **0.20** | 0.50 | 0.10 | **0.30** | **0.10** | **0.20** | **0.60** | 0.21 | **0.10** | 0.19 |

[1] http://dbpedia.org/Wiktionary

The results in Table 3.3 exhibit an upper bound with non-linear approaches given that particular combination of training and evaluation datasets. We also observe that performance depends on the density of cross-lingual connections among the language pairs, as well as their corpus size. Languages with smaller datasets and fewer alignments, such as Catalan, performed well when they were the source language, but worse when they were the target language. Some language pairs exhibit poor performance in both directions, suggesting that many of the tested words were not successfully captured by the alignment, regardless of the method, a possible consequence of inaccurate or too coarse alignments. While computationally expensive, CCAR is the best approach when translating rare terms between language pairs where a sufficiently good linear mapping suffices, like among Spanish and Catalan. Also, this particular language pair has denser cross-lingual links in Wikipedia.

Given our original hypotheses, we observe the following: while LSQ-SKIP shows more robust performance on all language pairs, their performance is not statistically significantly better than the baselines (outside of CLLSI, clearly outperformed at $p = 0.001$ on a paired T-test).

Within other experiments that have been omitted for brevity, we observe that by using other regularization strategies (such as L1 or Elastic Net) or loss functions (such as absolute or hinge loss) there are no statistically significant differences compared to the ones reported for LSQ-SKIP. We also observe that using CCA on top of embeddings (CCAT-SKIP) doesn't get the benefit of both: the embedding representation already implicitly models correlations between words of the same language, and CCA doesn not bring any additional improvement over simpler cross-lingual mapping methods. This suggests that if a pair of words correlates in one language, its counterpart will also tend to correlate in another. However, when observing non-linear approaches SIGLSQ-SKIP and EXT-SKIP, we observe that they are equal or better than the linear mapping based approaches. While EXT-SKIP does perform better on one language pair, there is no statistically distinguishable difference among them. However, the price in computational requirements is higher: for instance, on the German-English pair with 250 latent components, SIGLSQ-SKIP uses 277 seconds for training, while EXT-SKIP uses 9168 seconds.

In order to have a wider coverage of language pairs that did not have sufficient aligned pairs in Wiktionary, we synthesized a set of bilingual word dictionaries using the Microsoft Translator service [2], consisting of 2000 words for every language pair. Since we can't consider the output of this system as a ground truth, we rely on the fact that it's an MT system that does not need to rely only on comparable corpora, and its output can represent a proxy benchmark for this experiment.

The precision-at-10 results in Table 3.4 show a similar picture to the Wiktionary experiment in 3.3. Many of the best-performing methods achieve very similar performance: except when translating from Spanish or Catalan, the non-linear methods all reach the same upper limit on the Microsoft Translator evaluation dataset. When having Slovene as the source language, the overall performance is relatively high: precision at 10 is up to 0.8 when targeting Spanish, 0.54 for when targeting Catalan, 0.60 when targeting German and 0.70 when targeting English. One explanation is that mapping from languages with a smaller corpus to one with a larger one is easier, since a bigger proportion of the documents in the source language have outgoing links. While Slovene is one of the smaller datasets (at 136313 documents), the same phenomenon appears also with Catalan as the source language, having 395185 documents.

Figure 3.9 visualizes the performance of these methods on various language pairs. It shows the asymmetry of the problem, showing that some languages are more difficult to

[2] https://www.microsoft.com/en-us/translator/business/translator-api/

Table 3.4: Precision-at-10 of dictionary induction of different approaches on different language pairs, measuring with Microsoft Translator Gold Standard.

| Method | es-ca | es-sl | es-de | es-en | ca-es | ca-sl | ca-de | ca-en | sl-es | sl-ca |
|---|---|---|---|---|---|---|---|---|---|---|
| CCAR | 0.00 | 0.01 | 0.04 | 0.05 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.00 |
| CLLSI | 0.12 | 0.08 | 0.08 | 0.09 | 0.14 | 0.03 | 0.08 | 0.08 | 0.37 | 0.27 |
| LRCCA | 0.28 | **0.17** | **0.12** | **0.19** | 0.59 | **0.10** | 0.25 | 0.19 | 0.68 | **0.54** |
| LSQ-SKIP | 0.22 | 0.09 | 0.10 | 0.09 | **0.63** | 0.07 | 0.27 | 0.09 | **0.79** | 0.41 |
| CCAT-SKIP | 0.03 | 0.01 | 0.06 | 0.05 | 0.33 | 0.01 | 0.07 | 0.05 | 0.60 | 0.19 |
| SIGLSQ-SKIP | **0.30** | 0.10 | 0.10 | 0.10 | 0.60 | **0.10** | **0.30** | **0.20** | **0.80** | 0.50 |
| EXT-SKIP | **0.30** | 0.10 | 0.10 | 0.18 | 0.60 | **0.10** | **0.30** | 0.15 | **0.80** | 0.49 |

| Method | sl-de | sl-en | de-es | de-ca | de-sl | de-en | en-es | en-ca | en-sl | en-de |
|---|---|---|---|---|---|---|---|---|---|---|
| CCAR | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | 0.04 | 0.02 | 0.01 | 0.02 |
| CLLSI | 0.27 | 0.25 | 0.12 | 0.07 | 0.08 | 0.10 | 0.12 | 0.05 | 0.06 | 0.06 |
| LRCCA | 0.47 | 0.48 | 0.57 | 0.18 | **0.18** | **0.22** | 0.38 | **0.13** | **0.10** | **0.10** |
| LSQ-SKIP | 0.50 | 0.69 | 0.56 | 0.14 | 0.10 | 0.07 | 0.39 | 0.09 | 0.07 | 0.07 |
| CCAT-SKIP | 0.21 | 0.38 | 0.25 | 0.03 | 0.03 | 0.04 | 0.19 | 0.02 | 0.01 | 0.02 |
| SIGLSQ-SKIP | **0.60** | **0.70** | **0.60** | 0.10 | 0.10 | 0.10 | **0.40** | 0.10 | **0.10** | **0.10** |
| EXT-SKIP | **0.60** | **0.70** | **0.60** | 0.11 | 0.10 | 0.10 | **0.40** | 0.10 | **0.10** | **0.10** |

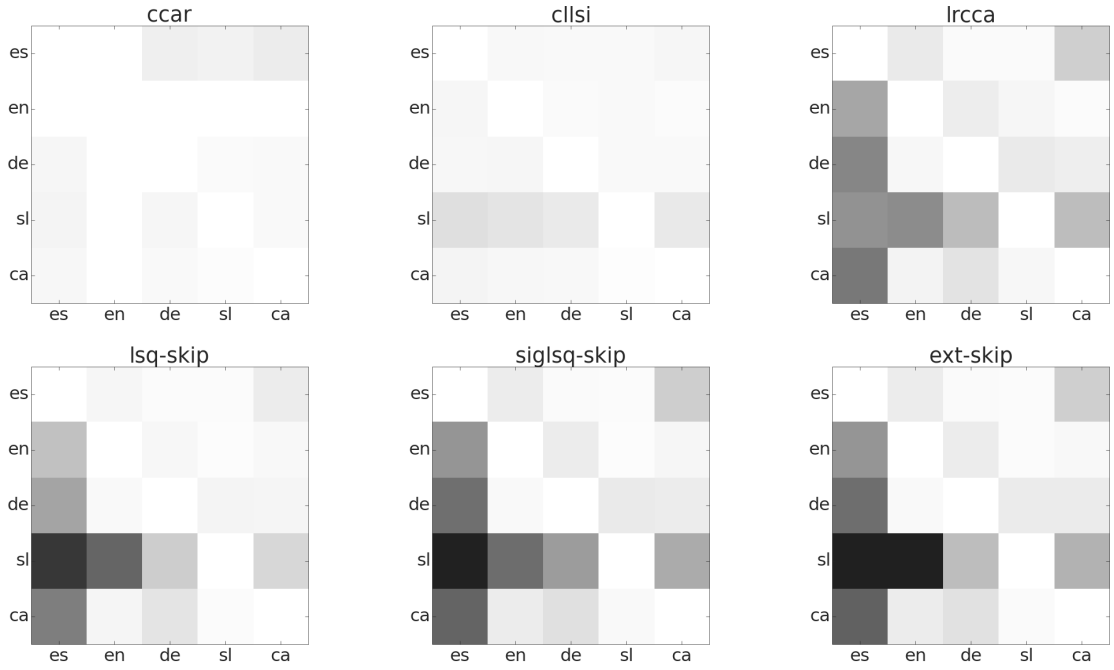


Figure 3.8: Visualization of results for every language pair given their Mean Reciprocal Rank score. The vertical axis is the source language, and the horizontal axis represents the target language. Darker tones correspond to better performance.

predict, and that languages with fewer documents are relatively easy to map to languages with larger corpora, especially where the cross-lingual linkage in the training dataset is good. Also, some languages are more difficult to predict due to their syntactic complexity, which makes the task of prediction of literal words less comparable across different language pairs. The performance could be improved by applying lemmatization to the individual words, especially since that would reduce the dimensionality of the space. The visualization also shows the difference that non-linear approaches have on the performance: the lower rows is visually notably darker than the upper row.

From the results, we can see that adding regularization to the solver via ridge regression alone improves performance on similar languages compared to a straight least-squares solution. We can improve the performance on less similar languages by employing a non-linear model in the pipeline.

#### 3.2.4.3 Experiments on cross-lingual similarity estimation

Since tasks involving multiple languages often involve operating with a similarity metric that works across languages, we also evaluate our methods on the task of approximating cross-lingual similarity. Cross-lingual similarity is a key feature in use cases such as linking news stories across different languages [89]. For this purpose, we use a comparable corpus across multiple language pairs, annotated with human judgments about the similarity of a given document pair [90]. Recent work on estimating similarity metrics using this particular dataset [91] shows that Wikipedia-specific metrics that also use the link structure perform best. However, they do not generalize to cross-lingual similarity between two arbitrary documents. The authors evaluate their performance by observing the correlation between their approaches and the actual judgment scores. The relevant MT-based approaches in the paper that are comparable to our setup are reported to obtain Spearman-rank correlations of $\rho_{de-en} = 0.47$ and $\rho_{hr-en} = 0.48$.

The dataset has three language pairs that can be tested by our infrastructure: German-English, Croatian-English and Slovene-English. Every language pair has 100 documents, each having two judgments. Within every judgment, there are four questions, the relevant one being the assessment on how similar a given pair of documents is on a five point scale.

We evaluate the approaches, presented in the previous sections on the task of estimating cross-lingual similarity. The goal is to produce a metric that correlates best with human judgments. For this purpose, we take evaluation document pairs $x_a, x_b$ from the test set, and translate them into $x_a b = f_a b(x_a)$ and $x_b a = f_b a(x_b)$, respectively, where $f$ is the translation method. Since we can measure similarity in both spaces, we define our similarity metric as the average of both mean absolute errors between the translations and the original:

$$\text{similarity}(x_a, x_b) = \frac{\frac{|x_a - f_{b\to a}(x_a)|}{|x_a|} + \frac{|x_b - f_{a\to b}(x_b)|}{|x_b|}}{2} \tag{3.14}$$

In Table 3.5, we observe that a similar ranking among methods holds also for the task of similarity estimation. While the Croatian-English similarity estimates correlate less than the MT-based system [91] (0.25 versus 0.48 Spearman correlation), the German-English ones correlate better (0.53 compared to 0.47). While the difference between the approaches can be explained by different translation training data, it also suggests that having a comparable corpus of a sufficient size (such as the German-English Wikipedia mapping) can perform even better than the $\text{trans}_2$ system that uses a MT system when estimating cross-lingual document similarity.

On the other hand, while we observe that all the approaches using word embeddings represent similarity better than the baselines, there is no significant difference among dif-

Table 3.5: Spearman and Pearsons's correlation coefficients between the predictions and ground truth similarity judgments.

| | $\rho_{\text{Spearman}}$ | | | $\rho_{\text{Pearson}}$ | | |
|---|---|---|---|---|---|---|
| Method | $de-en$ | $hr-en$ | $sl-en$ | $de-en$ | $hr-en$ | $sl-en$ |
| $\text{trans}_2$ [91] | 0.47 | 0.48 | | | | |
| CCAR | 0.516 | 0.196 | 0.446 | 0.467 | 0.133 | **0.361** |
| CLLSI | 0.506 | 0.228 | 0.440 | 0.472 | **0.174** | 0.355 |
| LRCCA | 0.471 | **0.247** | 0.362 | 0.452 | 0.167 | 0.338 |
| LSQ-SKIP | 0.533 | 0.204 | **0.450** | 0.472 | 0.158 | 0.355 |
| SIGLSQ-SKIP | 0.526 | 0.206 | **0.450** | **0.474** | 0.159 | 0.356 |
| EXT-SKIP | **0.537** | 0.208 | **0.450** | 0.473 | 0.159 | 0.356 |

ferent linear and non-linear models for translating among word embeddings, suggesting that estimating similarity is a simpler task than translation. Thus, we conclude that for estimating similarity, there is little need to employ non-linear estimators, with both the Spearman and Pearson correlation scores pointing to the same conclusion.

### 3.2.5 Discussion

In this section, we proposed a novel cross-lingual representation learning method for using comparable corpora that uses word embeddings words combined with a kernel approximation mapping that outperforms several baselines while keeping its scalability characteristics.

We show that for the task of dictionary induction, we often outperform linear approaches for translating across word embeddings. The approximate methods reach the same level of performance than regression tree translators. Given the results, the most efficient method among the dominating group is SIGLSQ-SKIP, having predictable learning times that are linear with regard to the number of data points, performing comparably to the computationally more expensive EXT-SKIP.

We also show that SIGLSQ-SKIP performed comparably to the computationally more expensive EXT-SKIP. None of the approaches among those using word embedding approach proved statistically significantly better than others, but overall, they still outperform the linear approaches on a majority of language pairs. A notable exception is the Catalan-Spanish language pair, where the CCAR proved to be highly precise when measuring with the Wiktionary bilingual dictionary due to the high relative connection density among that language pair, and the ability of CCAR to translate rare words.

We show that given the additional constraint of having parallel corpora, may still get good dictionary induction, especially where the target language has a sufficiently rich monolingual corpus to train word embeddings on.

We have also demonstrated the presented approaches on the task of similarity estimation. While all approaches using word embeddings outperform baseline approaches, we do not observe comparative benefits when using non-linear mappings to estimate similarity. The results demonstrate that for the similarity estimation task, such a system trained on a comparable corpus can perform at an equal level of performance to a machine translation system provided that the alignments in the comparable corpus are of sufficient density.

This work evaluated Wikipedia as a document-aligned data source. While its quality was sufficient to demonstrate algorithm improvements, the overall results were still poor. Refining the alignment to consider document sections and even individual sentences would likely improve the results.

## 3.3 Sparsity in Cross-Lingual Translation

This subsection describes a novel approach for translating vector space text representations from one language to another using only comparable bilingual corpora, focusing on the sparsity aspect of the problem. In this context, sparsity refers to property of the resulting vector to have a low-number of non-zero dimensions. Whether achieving sparsity is desirable at all, depends on the task at hand. For instance, typical information retrieval tasks can potentially show higher performance when using query expansion, which is analogous to producing a denser result. On the other hand, for tasks like bilingual lexicon estimation, the goal is to produce a translation using only a few salient terms. Extending this to the general case of vector space translation, we expect most of the information of the translated document vector to be concentrated on a few non-zero dimensions, comparable to the shape of the input vector.

The problem of translating across languages in their vector spaces is related to the task of bilingual lexicon estimation, where we generate a lexicon mapping from corpora in both languages. While most approaches that target this are using parallel corpora [92], or extract sentence pairs from documents in comparable corpora [93], this section proposes a method that produces vector space translations of documents without the assumption of alignment of individual sentence pairs.

This section compares against existing approaches that are typically used in cross-lingual information retrieval to produce such projections. This also motivates our use of the Mean Reciprocal Rank (MRR) as an evaluation measure, a metric that captures how highly the correct result ranks among all other options.

However, none of the baselines focused on information-retrieval are designed with sparsity of the output in mind. In other words, the methods discussed in Subsection 3.2.3 transform a sparse input example $x \in X$ into a dense projection.

While any method can be modified to produce a sparse output by setting the dimensions below a certain threshold to zero, it may not give an optimal solution under these sparsity constraints. Since the data in the input (the cross-lingual query in the retrieval scenario) is natural language in a bag-of-words vector space model, intuition suggests that its projection in another language should also be a sparse vector.

We present an approach based on translating documents through a low-dimensional common space provided by Canonical Correlation Analysis, posed as an optimization problem with sparsity constraints and evaluate its performance on vector space translation as a multivariate regression task.

The scientific contribution of this section is a method that improves the model fit of vector space translation using comparable corpora by applying sparsity constraints using an Elastic net solver. We show its effectiveness in several language pairs. We also evaluate the effect of enforcing sparsity, showing improvement on both bilingual mate retrieval and translation correspondence.

An area of research that also shares the problem of sparse reconstruction is compressive sensing [94], a signal processing technique that is able to efficiently acquire and reconstruct a signal by finding solutions to underdetermined linear systems. The common ground between compressive sensing and vector space translation is that we can look at the pair of two languages as a pair of a signal and its measurement, where we are trying to reconstruct the original signal given that we can define its transform, inverse transform and the measurement. These similarities in frameworks have already been used in applying sparsity in natural language processing [95].

When describing our domain in signal processing terminology, the measurement is the source language content, the transform is the translation component, and the signal that we are reconstructing is the content in the target language. The assumptions between the

domains of signal and natural language processing are analogous: compressed sensing also assumes that the signal can be represented with a low number of non-zero dimensions. This assumption corresponds to our assumption that the target translation should have a non-zero number of words that correspond to that of the source content.

### 3.3.1 Problem formulation

We continue with the same notation as in Section 3.2. We postulate that if an example $x_* \in X$ is sparse, its corresponding $\Psi(x_*)$ should also be sparse, having a better correspondence to its aligned example $y_* \in Y$.

While the output of any method can be trivially made sparse by simply trimming its coefficients to zero below a certain threshold, this does not provide an optimal solution. This is most obvious in the case of synonyms that typically appear in similar contexts. A translation approach that would take into account sparsity should pick only the highest scoring of the synonyms instead of all of them, leaving more room for other information. While typically producing a set of synonyms would be preferable to just one in many scenarios, we wish to optimize for the case where we wish to maintain the sparsity of the result.

We treat this problem setting as a case of multivariate regression, trying to predict a given $y_*$ given $x_*$. Given the nature of the problem, we use the coefficient of determination ($R^2$) as the evaluation metric to measure the quality of fit. $R^2$ tells us how much of the variance of $X$ can be explained by $Y$. While the maximum value of 1.0 implies total correlation between $X$ and $Y$, the values can also be negative, a result of having random correlation. Intuitively, $R^2$ corresponds to how well the individual words in the projected examples correlate with their true distributions in the test set that measures the correctness of the bilingual lexicon estimate. Its best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). For example, a constant model that always predicts the expected value of $y$, disregarding the input features, would get a $R^2$ score of 0.0.

If $\hat{y}_i$ is the predicted value of the $i$-th sample and $y_i$ is the corresponding true value, then the score $R^2$ estimated over $n_{\text{samples}}$ is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2} \tag{3.15}$$

where $\bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i$.

To get a better intuition about how sparsity affects related tasks, we also evaluate the approach on bilingual mate retrieval. Given that we have a matched pair $(x^*, y^*)$, we calculate the cosine similarities between $\Psi(x^*)$ and every $y \in Y$. The mean reciprocal rank (MRR) value is then the inverse of the rank of $y^*$ in the similarity ranking.

### 3.3.2 Elastic CCA

We propose a novel method to solve the sparse CCA projection problem. The method is called Low-rank CCA translation with an Elastic net penalty, Elastic CCA in short. It exhibits intrinsically sparse outputs of $\Psi(x_i)$ without requiring a separate post-processing step. Instead of using the explicit solution from LR-CCA, we instead pose the relationship as an optimization problem and apply a solver that also applies a $l_1$ penalty to the solution, leading to sparse coefficients. However, we are not interested in sparse coefficients in the original sense, but a sparse result that requires a different formulation. For this purpose we use the Elastic net regularization that minimizes the least-squares solution to a problem,

regularized by a linear combination of $l_1$ and $l_2$ penalties. Let $V$, $U$, $A$ and $B$ be the intermediate mappings for the LR-CCA solution to the mapping between $X$ and $Y$, $\alpha$ be the regularization parameter, and $\rho$ the balancing coefficient between $l_1$ and $l_2$. The CCA-ELASTIC poses the following optimization problem, where $\alpha$ and $\rho$ denote the regularization hyperparameters:

$$\Psi_{cca-elastic}(x^*) = \arg\min_{y^*} |x^* V^T A - y^* U B| + \alpha\rho|y^*|_1 + \frac{\alpha(1-\rho)}{2}|y^*|_2 \tag{3.16}$$

This approach satisfies the translation sparsity property that we require in our use case due to the dual regularizers: while we use the $l_1$ regularization for its sparsity effect, adding the $l_2$ makes the optimization problem convex, and easier to solve [73].

Intuitively, the solution to our approach would be the minimal set of words in the target language that would result in the same common space mapping as the words in the source language.

To evaluate whether sparsifying makes sense for other approaches and to provide comparable baselines, we use an adaptive approach to hard thresholding, denoted by -AHT, where we keep the top $k$ dimensions by weight, setting the rest to zero. We determine $k$ to be the same as the dimensionality of the input document. The intuition behind that is that document with $k$ non-zero words in one language translates to roughly $k$ non-zero words in the target language. We also test its variation, where we add a factor for the desired dimensionality: -AHT-2 would map $k$ non-zero dimensions from one language to $2k$ non-zero dimensions of the target document.

### 3.3.3 Experiments

Besides using $R^2$ as an evaluation metric for the prediction quality, we also measure the mean reciprocal rank (MRR) in mate retrieval, the task of trying to guess the correct corresponding data point in a parallel or comparable corpus given its counterpart. Having a matched pair $(x*, y*)$, we calculate the cosine similarities between $\Psi(x*)$ and $y \in Y$. The MRR value is then the inverse of the rank of $y*$ in the similarity ranking. While this metric is more suited for measuring the performance from the perspective of information retrieval, we report it for illustrative purposes.

To validate the models, we have chosen a set of 200 random documents pairs from Wikipedia as test sets, and the remainder as training sets. we tuned the model hyper-parameters using a grid search over the hyper-parameter space.

The comparable corpus used in this evaluation is Wikipedia in multiple languages, as described in Subsection 3.1.3. We use the subset of 65079 documents that are present in all seven languages. We transform the text into a bag-of-words representation with the TF-IDF weighing scheme and normalize the individual document vectors.

We treat this evaluation as multivariate regression, trying to predict $Y$ given $X$ as inputs, measuring the model fit using the $R^2$ metric.

We compare our proposed method, ElasticCCA with several baselines: CCA Regression (CCAR), as defined in Subsection 3.2.3.3, Low-rank CCA translation (LR-CCA), as defined in Subsection 3.2.3.2, and Cross-lingual LSI (CL-LSI), as defined in Subsection 3.2.3.1, as well as their variants with hard thresholding or adaptive hard thresholding post-processing, denoted with -HT.

We attempt three different approaches with hard thresholding. The first approach is -HT-K, where we keep the top $k$ dimensions, sorted by their weight. The second approach is adaptive hard thresholding -AHT, where we set $k$ to be the same as the dimensionality of the current source document.

This subsection summarizes the effects of the representation dimensionality and the different discussed thresholding methods. For LR-CCA and CL-LSI, we vary the number of latent vectors, denoting them with LR-CCA-$n$ and CL-LSI-$n$, respectively. For ElasticCCA, we vary both the number of latent vectors $n$, as well as the Elastic Net regularization parameter $\alpha$, denoted with ElasticCCA-$n$-$\alpha$. Wherever applicable, we also vary the $-HT$ options for the extent of hard thresholding.

We tuned the hyperparameters of the tested methods using grid search over the $\alpha$ parameter space and observe the results on the tuning set. For CCAR, we vary the regularization parameter from 0 to 1, showing an optimum at a mid point. One observation from results of hyperparameter tuning is that increasing the dimensionality does not help beyond 300 dimensions of the representation space in none of the methods.

Table 3.6: Effect of thresholding strategies on performance.

| Method | $R^2$ score | MRR |
|---|---|---|
| Thresholding strategies of CCAR | | |
| CCAR-0.55-HT-1000 | -17.48 | 0.1154 |
| CCAR-0.55-AHT-1 | **-2.66** | **0.1212** |
| CCAR-0.55-AHT-2 | -4.03 | 0.1186 |
| Thresholding strategies of Low-Rank CCA | | |
| LR-CCA-300-HT-1000 | 0.45 | 0.0548 |
| LR-CCA-300-AHT-1 | **0.54** | **0.0576** |
| LR-CCA-300-AHT-2 | 0.43 | 0.0566 |
| Thresholding strategies of Cross-lingual LSI | | |
| CL-LSI-300-HT-1000 | 0.54 | 0.0607 |
| CL-LSI-300-AHT-1 | **0.56** | **0.0626** |
| CL-LSI-300-AHT-2 | 0.50 | 0.0619 |
| Sparse result via ElasticCCA | | |
| ElasticCCA | **0.62** | 0.0756 |

The next question relates to the thresholding strategies. Results in Table 3.7 show that in the baseline methods, such as LR-CCA, CL-LSI and CCAR, adaptive hard thresholding that gives translations that have the same non-zero elements that its sources produce the best results, suggesting that the extra non-zero values present in AHT-2 do not bring additional information. Across all models, adaptive thresholding performs equivalently or better than fixed dimensionality thresholding.

Table 3.7: Average $R^2$ and MRR results across all language pairs for both thresholded and non-thresholded variant of models, as well as the natively sparse output of ElasticCCA.

| Method | $R^2$ | MRR |
|---|---|---|
| CCAR-0.55 | -36.79 | 0.1085 |
| CCAR-0.55-AHT-1 | -3.99 | **0.1222** |
| LR-CCA-300 | -0.14 | 0.0519 |
| LR-CCA-300-AHT-1 | 0.41 | 0.0544 |
| CLLSI-300 | -0.03 | 0.0600 |
| CLLSI-300-AHT-1 | 0.46 | 0.0705 |
| ElasticCCA | **0.62** | 0.0756 |

However, ElasticCCA obtains the highest overall $R^2$ scores. Furthermore, compared

to other methods based on low-rank approximation of the vector space, it also seems to improve when increasing the dimensionality of the low-rank space.

Averaging across all languages, results show that removing low-weight dimensions is critical for achieving high $R^2$ coefficients. In the majority of method pairs, the hard thresholding of coefficients significantly improved the performance over the dense result. On the other hand, the inherently sparse result of ElasticCCA achieves even higher performance of 0.62 $R^2$, averaging across all languages, ranging from 0.34 at en-zh to 0.68 at en-hr.

### 3.3.4 Comparison across language pairs

While the previous subsection focused on the overall performance, we also consider the performance of cross-lingual vector space translation in different language pairs.

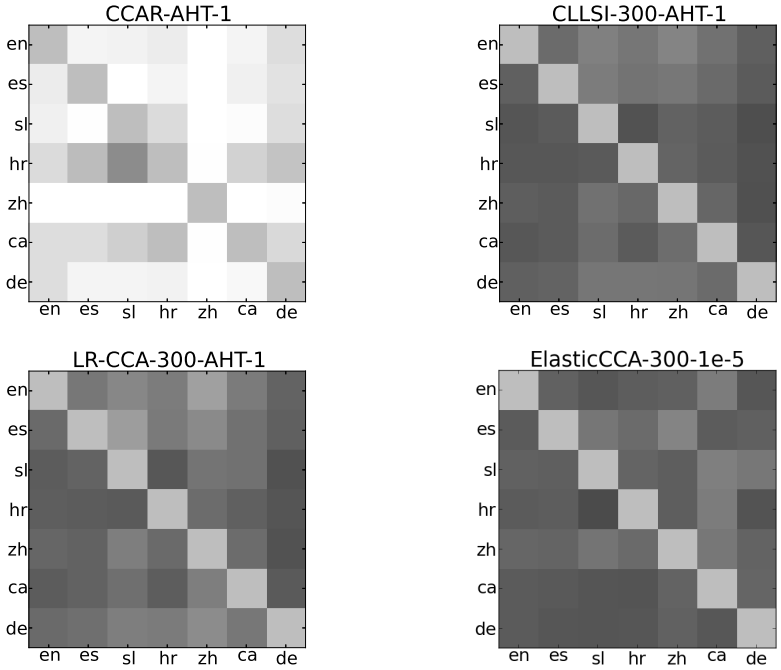


Figure 3.9: Visualization of results for every language pair given their $R^2$ coefficients. The luminance is an exponential transform of $R^2$, darker tones corresponding to better performance.

The visualization of the results in Figure 3.9 shows the $R^2$ performance on the best-performing versions of all baselines, along with two variants of regularization for ElasticCCA. Improvements shown in aggregate in 3.7 are consistent across language pairs. While the rows represent the source language, the columns represent the target language. The grayscale luminance represents how well the obtained translations correlate with their gold standard mates: pitch black corresponds to $R^2 = 1.0$, while white corresponds to $R^2 = -\inf$. Mid-gray values shown in the diagonals correspond to $R^2 = 0.0$. However, the visualizations reveal that some methods operate better on some language pairs: while CL-LSI and LR-CCA have an edge when translating to German, ElasticCCA has overall higher scores on other language pairs.

The visualization also suggests that for a given language pair, translating in one way can be easier than the reverse: for instance, translating to Spanish generally works better than translating from Spanish.

In terms of running time, the methods have different profiles. CCAR is a query-time

process, having an iterative solver, every iteration containing several operations that involve the entire corpora matrices, making it the slowest at the translation time. Training CLLSI involves a decomposition of the bilingual corpus matrix, where approximate approaches can make the runtime linearly dependent on the size of the corpora. The translation itself is fast, since it operates on the low-dimensional representation. LR-CCA has a similar front-loaded profile: training involves a SVD decomposition of the cross-covariance matrix, followed by a CCA of the transformed inputs. ElasticCCA has the same training as LR-CCA, but the translation uses an Elastic net solver for every query. Operating on a low-rank representation makes it faster than the same step in CCAR.

### 3.3.5 Summary

The results show that the ElasticCCA method, designed with sparsity in mind, outperforms methods that use adaptive hard thresholding as a post-processing step on most language pairs, with the exception when translating to German, while CCAR still maintains the best mate retrieval performance.

## 3.4 Named Entity Disambiguation

In this section, our goal is to explore possible improvements of named entity disambiguation (NED) quality by using ontologies in different ways along with statistical knowledge. To achieve this, we experiment with using different kinds of available data that could help in improving in-text entity disambiguation quality. Since entities that are related tend to appear together in documents more often, we explore the possibilities of expressing relatedness in different ways, such as similarities of entities' descriptions, the entity graph topology, and entity co-occurrence information.

Integration and sharing of data across different data sources is the basis for an intelligent and efficient access to multiple heterogeneous resources. Since a lot of knowledge is present in plain text rather than a more explicit format, an interesting subset of this challenge is integrating texts with structured and semi-structured resources, such as ontologies. This is especially interesting in the context of Open Linked Data [96], where the main motivation is to have cross-dataset mappings across as many datasets as possible. However, textual datasets have to be treated differently in some ways. This involves dealing with natural language ambiguities in names of entities.

For example, in the case when we have a document where there are two unknown entities referred to by the names "Elvis" and "Memphis". The first is a common personal name and the second one the name of several locations. We would like to use this relatedness information between those two entities to help in resolving "Elvis" as a well-known singer and "Memphis" as a city in Tennessee, where the identified singer lived.

A long-term goal of this work is to improve the quality of in-text entity disambiguation using existing ontologies and mappings between them, as well as to reduce the language barrier when integrating text and knowledge. In other words, we would like to be able to bootstrap existing knowledge with the intention of obtaining new knowledge. Furthermore, named entities play a special role in multi-lingual use-cases: in machine translation [97], the translation mechanisms for named entities depend on both the source and target languages. There are specific rules to translate (or transliterate) particular proper names or concepts. In order to support this use case, we propose an automatic natural language processing method to annotate the content so that it can be correctly processed.

The work in this section has been published as Štajner and Mladenić [98].

### 3.4.1 Problem Definition

We state the problem as identifying and disambiguating concepts that appear as mentions within a fragment of text. Disambiguation is important because phrases may have many distinct meanings. While human readers are able to infer the meaning from context, this task is difficult for computers. For instance, the phrase "Washington" can be either a person, location or an organization, and even constraining its type to a location yields over sixty possible different locations having that name.

We use the ontology as background knowledge represented as a graph of entities, identified with URIs, described with attributes and interconnected with different relationships to other entities. Entities also need to have textual labels and descriptions necessary for entity disambiguation via vector space model similarity [38, 99, 100, 101].

Well-defined entities and relationships are a property of the knowledge model that asserts that a single term only has a single meaning. In that case, we refer to terms as entities. We achieve this property by performing entity disambiguation. In general, state-of-the-art entity disambiguation systems [38, 39] use heuristics that fall into three main categories.

- **Mention popularity** captures the overall most likely meanings of entity phrases. It is typically modeled by the conditional probability of the named entity given a mention.
- **Context similarity** captures the entity that best fits the topical context around the mention. It is modeled by the similarity of the mention's context and the entity's context, using a similarity measure operating on a bag-of-words model. The mention's context is a window of words around the mention in the input text, and the entity's context is its description.
- **Coherence** collectively captures the entities that make sense appearing together because they are somehow related to one another. While context similarity operates on a single mention-entity pair, the coherence heuristic is collective, operating on the whole input document. It is typically solved by a greedy graph pruning algorithm.

Named entity extraction is the task of using the surrounding context to isolate the part of text which represents an entity, referred to by a proper name. It is often coupled with entity classification, determining to what class it belongs to, for instance a person or an organization. In general, these are implemented as supervised sequence classifiers. For this purpose we used the Stanford Named Entity Recognizer [102].

Before using our background knowledge base, we can still perform a part of co-reference resolution with the identified entities, such as canonicalization, partial name consolidation and acronym consolidation. Simple de-duplication of extracted entities also helps to reduce the search space when performing collective entity disambiguation. Once we have a basic understanding of which distinct entities we are trying to resolve, we can search our knowledge base for possible candidates that could match the named entities. We then perform a series of decisions, where entities from the document are matched with the most relevant knowledge base entity based on some relevance criteria. This is then repeated as long as there are unmatched entities in the documents or none of the remaining candidates fulfill the minimum criteria for matching.

This section also provides means to challenge the assumption that the language of the input document is the same as the language of the knowledge base, which is not always the case.

### 3.4.2 Using entity relatedness as a coherence estimate

When solving the entity disambiguation problem, the usual approach involves performing graph clustering over the entity graph using a certain similarity criterion [43]. In context of relational data, it is a combination of attribute similarity and relational similarity. However, such approaches are more often found in structured data, whereas our approach attempts to use these techniques on linking unstructured text with semi-structured data. Also, when using ontologies as a sense inventory, relationships between entities are heterogeneous, meaning that not every relationship is equally informative, especially since a large portion of relationships in a typical knowledge base is simply declaring the type classes of various entities. The proposed method solves this problem by introducing a relatedness metric that is aware of the difference in informativeness. The authors make a distinction with regard to the independence assumptions of the disambiguation operation:

- **Pair-wise disambiguation**: decisions are done independently for each mention of an entity in the document
- **Collective disambiguation**: decisions do not assume independence of disambiguation decisions, enabling us to use relatedness data in the subsequent decisions.

Since collective entity disambiguation can take relatedness between entities into account to estimate the coherence of a given solution, we experiment with the following definitions of relatedness metrics:

- **Content similarity** as a relatedness measure is relevant in situations where only available data is in form of attributes and textual descriptions and no explicit relationships between entities. This is the typical approach that is used in document clustering, which operates under the assumption that textual documents that are related or part of the same topic also have higher content similarity among themselves.

- **Entity co-occurrences** as a relatedness measure are useful in situations where we can obtain a training corpus of documents, annotated with resolved in-text entities. In general, co-occurrences are a common source of training data for information retrieval problems, analogous to entity disambiguation. Use cases that apply this technique range from protein identification [103] to resolving geographical locations [104]. Co-occurrences have also been used to construct a generative model [36] for entity disambiguation.

- **Explicit relationships** as a relatedness measure: relationships between entities are the most explicit form of relatedness. However, not all relationships have the same significance. This section proposes one such possible approach to heterogeneous relational entity disambiguation that bases relational significance on the frequency of the relation appearing in the knowledge base with regard to entity types. This measure was suggested [42] as one of the suggested methods of determining a minimal informative sub-graph of a graph. Since this problem as well as multi-relational entity disambiguation both use the notion of relational significance, this section will explore the possibilities of using this measure as a means of quantifying relatedness between entities.

#### 3.4.2.1 Independent entity disambiguation baseline

When matching an entity from the document to a candidate entity, we employ some heuristics to evaluate the confidence of their match. One such heuristic is description similarity. Note that since this scenario has no a priori matches of document entities with knowledge base entities, we have no use for relational information.

When a single entity appears in multiple documents, our task is to evaluate each candidate and finally match the document entity with the top candidate. Description similarity is defined as the cosine similarity of TF-IDF vectors of descriptions that represent the given entities. Since one of the entities is a document entity, its descriptions is the document text itself. We then resolve each entity in the document to its most similar candidate among the candidates from the knowledge base.

#### 3.4.2.2 Collective resolution with relatedness

While leaving behind assumptions of independence, we can then benefit from using information on relatedness between entities. Collective candidate selection is performed with the following sequence of steps, adapted from the relational clustering algorithm for entity resolution [43] and adapted from the general dataset reconciliation domain to a text-knowledge base alignment scenario.

Algorithm 3.3 describes the adapted entity disambiguation algorithm. The high-level operation is the same for all described approaches.

**ALGORITHM 3.3:** Collective entity resolution algorithm.

**Input**: document entities $D$, candidate entities $C$, priority queue $Q$
**Output**: selected matches $S$

```
S ← ∅
for potential pair between f ∈ D and candidate entity e ∈ C do
    Q.Insert(relevance_pairwise(f, e), f, e)
    while Q ≠ ∅ do
        relevance_{f,e}, f, e = Q.Pop()
        S.Add(relevance_{f,e}, f, e
        for q ∈ Q, q = (r, a, b), a = f do
            Q.Remove(q)
        end
        for (r, a, b) ∈ Q do
            r ← relevance_collective(a, b, S)
        end
    end
end
```

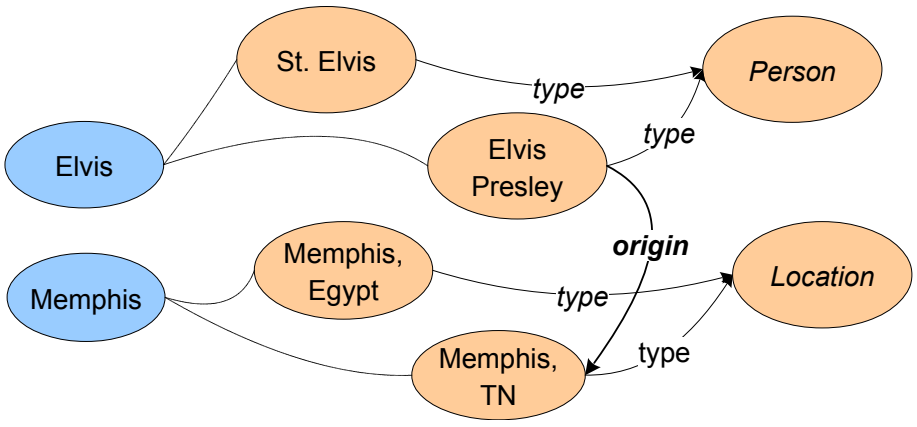


Figure 3.10: Using different semantic relations as a relatedness measure.

The three approaches differ only in the calculation details of the relatedness estimate that is used in collective relevance calculation, as defined here:

$$relevance_{collective}(f, e, S) = relevance_{pairwise} + \lambda \cdot \frac{\sum_{e_s \in S} related(e, e_s)}{|\{e_s \in S; related(e, e_s) \neq 0\}|} \quad (3.17)$$

The following subsections will describe the respective relatedness estimation approaches.

#### 3.4.2.3 Using semantic relations from the knowledge base

In Fig. 3.10, the blue nodes (Elvis and Memphis on the left) represent the document entities, whereas all the other nodes (colored pink) represent entities from the knowledge base. In this case, the relatedness between the entities is expressed explicitly in the form of RDF statements in the background knowledge - as shown in Figure 3.10. Consider the case where the subject *Elvis Presley* in relation *Hometown* (as his *origin*) to the subject *Memphis, Tennessee*. For use in our resolution model, we interpret relations as links with a specified weight. If the relations in the knowledge base are only of a single type, they are equivalent. However, when dealing with heterogeneous ontologies, as is often the case,

one has to estimate the importance of each link. For instance, if the knowledge base contained the RDF statement $\langle ElvisPresley, type, Person\rangle$, this would not be too useful, since it would likely encompass every entity called "Elvis" since they are mostly of the type *Person*. On the other hand, the relation $\langle x, Hometown, Memphis_Tennessee\rangle$ is a strong indicator, because it covers a much smaller set of entities. This property is defined as selectivity, and its value represents the cumulative weight of the entity's links. Determining the selectivity of the links is a problem, similar to finding the most informative sub-graph in a given semantic graph [42]. The authors wanted to find the smallest sub-graph, which would be sufficiently informative. For the purposes of determining subsets of the links they have developed multiple metrics to estimate the selectivity. One of the proposed metrics, which is also suitable for our domain, is Instance Participation Selectivity, which stipulates that the selectivity of the assertion $\langle s, p, o\rangle$ is inversely proportional to the number of statements RDF which correspond to the $\langle type(s), p, type(o)\rangle$ where the predicate $type(x)$ is defined as the relation of rdf:type of the entity. Let $\pi(type(s), p, type(o))$ be the set of all statements in the domain knowledge base, where the type of subject is $type(s)$, the predicate is p and type of object is $type(o)$.

$$IPS(s, p, o) = \frac{1}{|\pi(type(s), p, type(o))|} \tag{3.18}$$

To balance the estimate values for our use case, this work modifies the equation slightly to:

$$IPS_{log}(s, p, o) = \frac{1}{\log(1 + |\pi(type(s), p, type(o))|)} \tag{3.19}$$

The consequence is that the link type $\langle Person, Origin, Area\rangle$ is less selective than $\langle Person, Origin, City\rangle$, which is also what we want to model. This approach therefore enables us to quantify the relatedness of a pair of entities based on knowledge base data. The direct relatedness score is then calculated as:

$$\text{relatedness}_{direct}(e_i, e_j) = \frac{\sum_{\langle e_i, p, e_j\rangle \in KB} IPS_{log}(e_i, p, e_j)}{|\{\langle e_i, p, e_j\rangle \in KB\}|} \tag{3.20}$$

where we apply the log term to dampen the effect of more frequent relationships, similar to how TF-IDF applies the log to the inverse document frequency to reduce the impact of frequently occurring words.

However, when considering actual relatedness, we also take into account not only direct relations, but also indirect ones - the relations to entities that are in their common neighborhood. We define this as:

$$Nbr(e) = \{f; \text{relatedness}_{direct}(e, f) \neq \emptyset\} \tag{3.21}$$

$$Nbr(e_i, e_j) = Nbr(e_i) \cup Nbr(e_j) \tag{3.22}$$

We define indirect relatedness as an average of paths between both entities:

$$rel_{indirect}(e_i, e_j) = \frac{\sum_{f \in Nbf(e_i, e_j)} rel_{direct}(e_i, e_j)}{|Nbr(e_i, e_j)|} \tag{3.23}$$

We compute the final semantic relatedness score as a linear combination of direct and indirect relatedness:

$$rel_{knowledgebase}(e_i, e_j) = \lambda_1 rel_{direct}(e_i, e_j) + \lambda_2 rel_{indirect}(e_i, e_j) \tag{3.24}$$

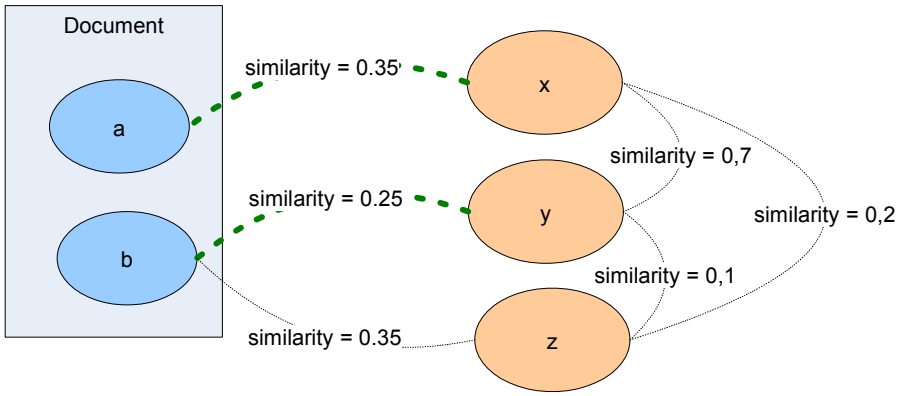


Figure 3.11: Using content similarity as a relatedness measure. The green dotted lines represent selected entities.

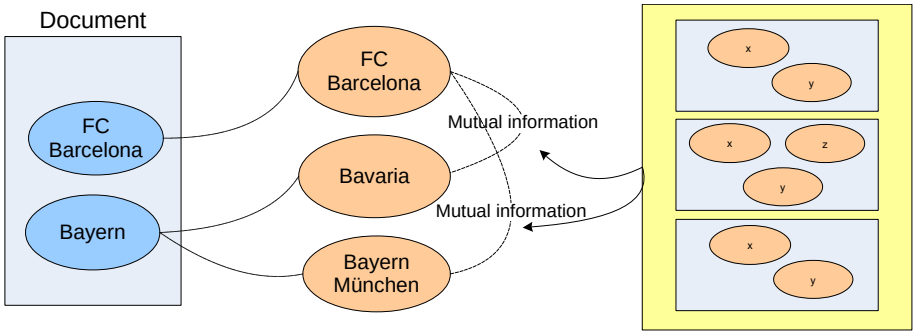


Figure 3.12: Using mutual information form entity co-occurrences as relatedness.

#### 3.4.2.4 Using content similarity

In some situations, we do not have explicit relations between entities. If the entities have descriptive attributes, we use them to estimate relatedness with comparing their content similarity, as illustrated in Figure 3.11. One advantage of such approach is that we do not require any more data than with pair-wise resolution, which adds to the flexibility of this method.

We formulate this as:

$$rel_{content}(e_i, e_j) = sim_{content}(e_i, e_j) \tag{3.25}$$

#### 3.4.2.5 Using co-occurrences from an annotated corpus

We can also represent relatedness between entities as co-occurrences, as shown in Figure 3.12. For example, the surface form "Bayern" can represent both the German federal state, or the football club ("FC Bayern") based there. When training on datasets containing sports articles, where multiple teams are mentioned, the mutual information between the meaning representing the team will be much higher than the meaning representing the region.

Data on the co-occurrences of two events are successfully used in information retrieval problems such as cross-language information retrieval [105] and determining the importance of words [106, 107], which is a problem related to entity disambiguation. Intuition for the use of the co-occurrences is the more often the two events occur together more frequently

than by chance, the more likely that they are related. This principle was also demonstrated with a collective generative model [38]. In our domain, we can model relatedness with point-wise mutual information [108] of two entities occurring in the same document.

$$rel_{co-occurences}(e_i, e_j) = PMI(e_i, e_j) = \log \frac{p(e_i, e_j)}{p(e_i)p(e_j)} \tag{3.26}$$

Since this procedure requires supervised learning, its output quality depends on the quality and coverage of the training corpus.

#### 3.4.2.6 Combining relatedness estimations

Since each relevance estimation method produces its own relevance score, it would make sense to have means of combining them. This is with expressing the relatedness function as a linear combination of all relatedness estimation functions:

$$rel(e, f) = \lambda_1 rel_{content}(e, f) + \lambda_2 rel_{co-occurences}(e, f) + \lambda_3 rel_{knowledgebase}(e, f) \tag{3.27}$$

The lambda parameters $\lambda_1, \lambda_2$ and $\lambda_3$ are experimentally obtained using a sequential Bayesian optimization with the Hyperopt library [109], using a separate development ground truth set.

#### 3.4.2.7 Evaluation

Our assumption is that the knowledge base consists of knowledge database that contains enough data to be able to perform the following tasks. First, it should be able to refer to each entity with multiple aliases to facilitate candidate retrieval. Second, it should be able to provide enough additional entity features, which we can use to compare those entities to each other and to article anchors that we attempt to link to.

Following these requirements, we chose to use a part of DBpedia, as described in Auer, Bizer, Kobilarov, *et al.* [110], since it provides both description and attribute data from Wikipedia, as well as references to other ontologies that describe other aspects of the same real-world objects. For the purpose of having rich heterogeneous relational data, we also used the Yago knowledge base, defined [111], which maps Wikipedia concepts to corresponding WordNet classes. Since a direct mapping from Yago to DBpedia exists, merging the two together is trivial. However, both ontologies are much broader than what our approach requires — we currently only use information on aliases, textual descriptions, rdf:type attributes and Yago categories of entities.

For determining the quality of the methods we have used precision and recall, measured at a certain level of confidence in the suggested entities for a given article. We then compared the suggested entities for those articles with manually identified entities of those articles. As a comparison measure, we use the area under the precision-recall curve.

Precision and recall are balanced with a relevance score threshold, selecting only those entities whose relevance score is above this threshold. This serves as a useful balancing tool, since in several examples, the entity does not exist in the knowledge base. In those cases, even the best candidate has a relatively low score.

We report the final results the value of $F_1$, which is the weighted harmonic average of precision and recall. Since we can apply a range of confidence thresholds to our system, we also obtain a range of $F_1$ values. Therefore, we report the area under that curve and the maximum $F_1$ as evaluation metrics.

We perform evaluation using the New York Times article corpus [112], using 39953 articles from January 2007 to April 2007 as training data for construction of TF-IDF weighted

Table 3.8: Performance of respective methods on entity resolution.

| Method | Area | max $F_1$ | Prec. max $F_1$ | Rec. at max $F_1$ |
|---|---|---|---|---|
| Collective | | | | |
| Relational, content, cooccurrences | 0.689 | 0.780 | 0.847 | 0.723 |
| Relational, cooccurrences | 0.689 | 0.780 | 0.847 | 0.723 |
| Relational | 0.683 | 0.778 | 0.849 | 0.718 |
| Cooccurrences | 0.679 | 0.773 | 0.839 | 0.717 |
| Content | 0.667 | 0.767 | 0.834 | 0.710 |
| Pairwise | 0.666 | 0.766 | 0.833 | 0.709 |

vectors. The articles were processed with an implementation of the described algorithm. For evaluating the performance of different approaches we manually selected and evaluated 945 entity disambiguation decisions from 79 articles as either correct or incorrect. Of the 79 articles, we used one third as a development set for hyperparameter tuning, while the remainder serves as the test set. Since the methods of pair-wise content comparison, collective content comparison and collective relational comparison are unsupervised, they do not require any labeled training articles. On the other hand, using co-occurrences as a relatedness measure requires training data. For this purpose, we take the remainder of the four months of articles that we did not use as a test set and process them with the collective relational comparison method. Since we wish to maximize the training data quality with our best effort, we use only entities whose relevance estimate is above a certain threshold. We used the threshold that gives us 95% precision using the development dataset. The collective relational comparison is used because it gives the highest quality output for this purpose. These values depend on a specific knowledge base and text corpus, so they are not necessarily universally applicable.

We perform the evaluation as an ablation study of different relational similarity metrics: this experiment starts with a system using all of the proposed collective similarity metrics, and shows whether removing one affects the performance. If removing a similarity metric term does not affect the performance, it suggests that term is not necessary to achieve the same performance. We compare collective disambiguation approaches to the pairwise disambiguation approach as the baseline.

Results in Table 3.8 show that additional information does indeed show improvement in $F_1$. The main metric that we use to compare approaches is the area over the precision-recall curve. As opposed to other reported metrics, it captures the performance over the entire precision-recall curve and not just a single point. The results show that removing the co-occurrence term reduces performance from 0.689 to 0.683. On the other hand, removing the content similarity term doesn't affect performance. Comparing the collective disambiguation approach using only relational similarity with the pairwise disambiguation alone, the difference is notable. While the same applies for entity co-occurrence similarity, collective disambiguation with content similarity performs roughly the same as standard pairwise disambiguation.

Figure 3.13 shows a clearer picture of the precision-recall trade-off when using different approaches. It demonstrates that the pairwise entity resolution approach (shown in purple) scores lower than any of the collective approaches across the entire precision-recall range.

As confirmed in Table 3.9, collective methods have an advantage over the baseline when looking at recall at 95% precision. Here we can also demonstrate improvement of using the combined model, yielding the best overall result. When calibrating for a given precision value, the differences in the obtained recall can be large when observing the recall at 95% precision.

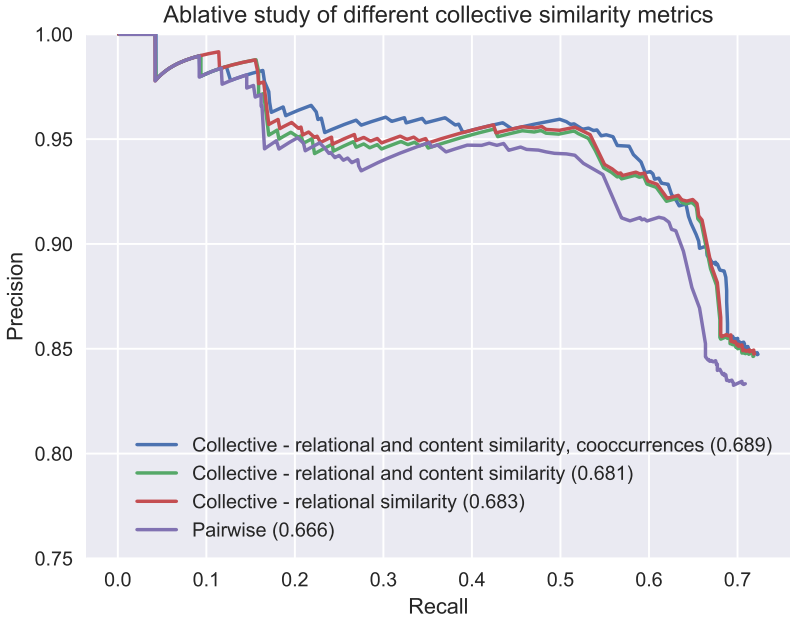


Figure 3.13: Comparison of entity disambiguation quality of different relatedness heuristics.

Table 3.9: Recall at two levels of precision: 90% and 95% precision.

| Method | Recall at 90% precision | Recall at 95% precision |
|---|---|---|
| Collective | | |
| Relational, content, cooccurrences | 0.656 | 0.559 |
| Relational, cooccurrences | 0.655 | 0.559 |
| Relational | 0.664 | 0.534 |
| Cooccurrences | 0.631 | 0.529 |
| Content | 0.630 | 0.222 |
| Pairwise | 0.630 | 0.203 |

#### 3.4.2.8 Discussion

When observing all the collective methods that we discuss in this section, the best performing ones were collective resolution with content similarity and collective resolution with relational weighing. The ablation experiment determined that the relational similarity and the co-occurrence terms were necessary to achieve the performance, while content similarity was redundant. One of the causes for this sort of behavior is that some documents tend to discuss unrelated entities. Furthermore, in longer texts, the entities, mentioned at the beginning of the document are not necessarily related to the ones on the other parts of the document, suggesting that we should experiment with taking the document paragraph structure into account.

Recent approaches, such as [113], resort to using neural networks architectures that explicitly model mentions, entities and their contexts. While fundamentally that approach uses the same data model as the pair-wise one described in this section, the neural network approach relyies less on hand-written heuristics that makes it easier to deploy in a production setting.

Using background knowledge as a disambiguation signal has also been explored in [114], where they introduce a selectional preference heuristic that tracks in what type of assertions does a given entity appear in, given the background knowledge. While the selectional preference heuristic plays a similar role to the relational selectivity heuristic described in this section, the actual implementation differs. Furthermore, the same paper mentions a distributional similarity heuristic, modeling a probability distribution of entities appearing together, which corresponds well to the collective resolution aspect described in this section.

### 3.4.3 Cross-lingual named entity disambiguation

When extending this pipeline into a scenario where the input and the knowledge base are represented in multiple languages, the biggest impact of this change is on the context similarity heuristic. Because it operates on the level of lexical similarity, its output has little meaning when the assumption of a single language is removed. The work in this subsection has been published as [115].

#### 3.4.3.1 Context similarity

We can approach the issue of language mismatch from several directions. A relatively simple approach to the problem is to see what kind of words carry contextual information.

The approach, presented in [44] assumes that most of the information is carried in other proper names. After testing several similarity measures, the authors report that a lot of the actual information is implicitly present in proper nouns, such as person and location names that appear on both sides, regardless of language. In practice, this reduces their similarity measure to something measuring entity overlap, since proper names are the only tokens that have any overlap in most language pairs: if the two texts share the same phrases, they are related. While this is true when the language is the same on both sides, it actually holds to some extent when the languages are not the same, as long as the nouns are not translated, transliterated, or orthographically modified in any other way. Based on this work, we use language-agnostic cosine similarity of two documents as a baseline context similarity method, denoted as *direct*.

Another baseline uses the equivalent of just using the context-independent 'mention popularity' measure, backed by a dictionary [45]. We construct the bilingual dictionary from anchor texts from non-English to English Wikipedia pages. While the authors show

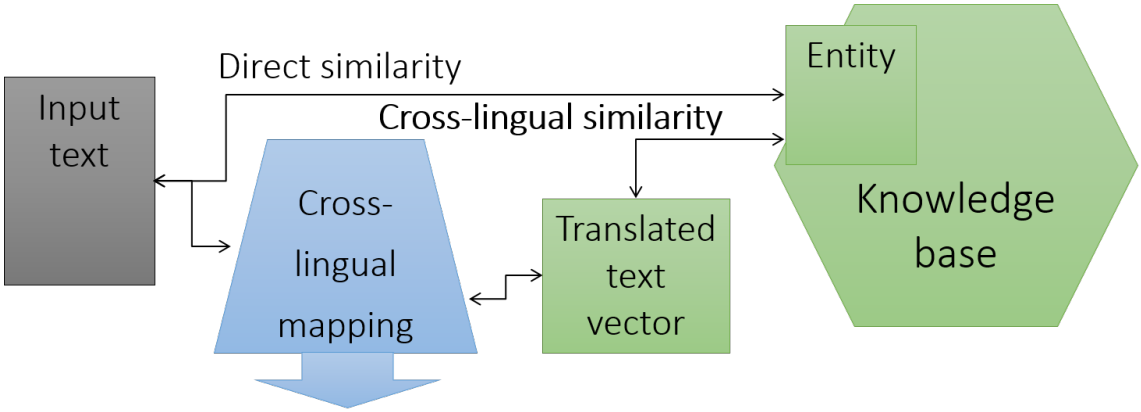


Figure 3.14: Architectures for cross-lingual named entity disambiguation

good results in estimating the prior popularity of an entity, it does not handle contextual cues.

A system representing an upper bound is the one that would actually translate the text of the document in the desired language and perform the disambiguation on the translation. While doing the translation manually is not feasible for our task, one may use machine translation to do this [46]. We implement an approach that uses Microsoft Translator to translate the input text and uses cosine similarity of the translated text as context similarity. We denote this approach with *machine translation*. While they achieve up to 94% performance of a monolingual baseline, machine translation greatly complicates and slows down the processing, opening a window for more efficient approaches.

We propose a method that incorporates a cross-lingual similarity measure into the framework. Instead of just computing literal context similarity between two contexts of different languages, we use an additional linear mapping that is able to map one vector of bag-of-words features into a vector in another language. This enables us to actually perform meaningful similarity computation on the same vector space.

The method used in this approach is Regression Canonical Correlation Analysis (CCAR), a dimensionality reduction technique operating on two views, finding a linear combination of vectors from both views (languages) that are maximally correlated [24]. The first vector corresponds to the input document, while the second one corresponds to the optimal mapping of it. However, instead of calculating this mapping in advance, we solve the optimization problem for each input document separately around the input document as the initial projection vector. A precise definition of CCAR is given in Subsection 3.2.3.3.

Figure 3.14 represents the two ways of obtaining a context similarity measure between an input document and one of the candidate entities. We can either do direct similarity, or we try to map the cross-lingual representation into a vector space, compatible with the knowledge base. While the former approach keeps the proper names intact, the latter may corrupt them: cross-lingual projections do not guarantee preservation of names.

Using a cross-lingual mapping exposes us to the risk of poor domain coverage. Initial experiments show that because the cross-lingual mapping was not able to map some words from the input document, it exhibited reduced performance. Therefore, we interpolate the cross-lingual similarity with the direct similarity with the proportion of the words that the cross-lingual mapping was able to recognize. We denote this approach as *CCAR-direct*.

In pre-processing, we use the Stanford Named Entity Recognizer [102] for English named entity recognition.

#### 3.4.3.2 Evaluation

Table 3.10: Performance of approaches on cross-lingual named entity disambiguation.

| Method | $F_1$ |
|---|---|
| Direct | 0.37 |
| Machine Translation | 0.46 |
| CCAR | 0.44 |
| CCAR-direct | 0.33 |

Experiments in Table 3.10 show that obtaining a cross-lingual mapping does improve on the context-similarity based NED when the training corpus and the input text share a common topic. Results show that the CCAR-based disambiguation context translation trails an MT-based context translation upper bound by a small amount (0.44 to 0.46).

Future work involves evaluating different cross-lingual similarity models, as well as transliteration models and data integration issues that arise when dealing with multilingual knowledge bases, as well as evaluating on other available datasets designed for this task, such as RECSA [116].

### 3.4.4 Named entity extraction for Slovene

When developing infrastructure for cross-lingual named entity disambiguation, entity extraction for Slovene was a missing piece of the puzzle. For this purpose, we developed and implemented a named entity extractor for Slovene language based on a machine learning approach [117]. It uses a CRF (Conditional random fields) model trained on the SSJ-500k corpus. Besides using several typical NER features, the recognizer is also enhanced with lexicons, morpho-syntactic tags, as well as trained on a longer feature window.

The corpus, available under a Creative Commons licence, is annotated with morpho-syntactic tags, as well as named entities for people, locations, organisations, and miscellaneous names. The approach uses morpho-syntactic tags, lexicons of entity types, and conjunctions of features of neighbouring words. Using all the best-performing features, the recognizer reaches a precision of 77% and a recall of 76%, having stronger performance on personal and geographical named entities, followed by organizations, but performs poorly on the miscellaneous entities, since this class is diverse and consequently difficult to predict.

Considering a comparable English and German named entity extractor using a similar implementation [118], and comparable datasets (CoNLL 2003), we can put the results for the Slovene NER in context. We obseve that while detecting person names is nearly as successful as in English, other entity classes pose a higher difficulty. We also see that the Miscellaneous class is much more difficult in the Slovene dataset compared to the other two languages.

A major contribution of the approach is also showing the benefits of splitting the class

Table 3.11: Comparing results on named entity extraction, measuring $F_1$ scores on multiple entity classes.

| Language | Person | Organization | Location | Misc. |
|---|---|---|---|---|
| Slovene [117] | 0.88 | 0.60 | 0.81 | 0.33 |
| English [118] | 0.91 | 0.79 | 0.87 | 0.72 |
| German [118] | 0.81 | 0.55 | 0.70 | 0.53 |

of miscellaneous entities into organizations and other entities, improving the performance even on personal and organizational names.

### 3.4.5 Discussion

To summarize, we decomposed the task of named entity disambiguation into three subtasks: estimating the prior distribution of entities given the mention, contextual similarity, and entity relatedness. We propose an approach for estimating coherence using three different ways of estimating relatedness among entities and show that it a combined approach outperforms a baseline that is just based on contextual similarity. When extending this problem to a cross-lingual situation, we needed to remove the assumption of a shared language. This affects the estimators for prior popularity and contextual similarity the most, which is why we propose an approach that employs a cross-lingual similarity model to capture the contextual similarity. We show that while cross-lingual disambiguation introduces engineering problems by itself, our approach offers a relatively simple solution using CCAR that rivals a much more complex proprietary machine translation engine, when used for cross-lingual similarity estimation.

The results of this research were used in the Enrycher system [119] and, the X-Like language processing infrastructure [120], as well as in the DiversiNews [121] system.

### 3.4.6 Applications in internationalization

This subsection describes how natural language processing techniques described in this chapter can improve the content localization process, a typical way of creating multilingual content. We describe the requirements from the localization side to support the use case of translating named entities and the resulting design constraints. Using this, we describe a standard data representation that was defined in the ITS 2.0 W3C standard [122], whose purpose is to standardize core concepts that are designed to foster the automated creation and processing of multilingual Web content. Having this capability allows integration of various language and localization tools into this process based on a common interface.

Content or software that is authored in one language (so-called source language) for one locale (e.g. the French-speaking part of Canada) is often made available in additional languages or adapted with regard to other cultural aspects. From the viewpoints of feasibility, cost, and efficiency, it is important that the original material is suitable for downstream phases such as translation. During the translation phase, the meaning of a source language text is analyzed, and a target language text that is equivalent in meaning is determined. For example national or international laws may regulate linguistic dimensions like mandatory terminology or standard phrases in order to promote or ensure a translation's fidelity.

The Internationalization Tag Set (ITS) 2.0 addresses some challenges and opportunities related to internationalization, translation, and localization. ITS 2.0 in particular contributes to concepts in the realm of metadata for internationalization, translation, and localization related to core Web technologies.

ITS 2.0 prescribes other data categories that are relevant in localization and content authoring workflows:

- *Translate*: expresses information about whether the content of an element or attribute is intended for translation or not.
- *Localization Note*: used to communicate notes to localizers about a particular item of content.
- *Terminology*: used to mark terms and optionally associate them with information, such as definitions.
- *Directionality*: allows the user to specify the base writing direction.
- *Language information*: used to express the language of a given piece of content.
- *Elements within text*: reveals if and how an element affects the way text content behaves from a linguistic viewpoint, for example to provide basic text segmentation hints for tools such as translation memory systems.
- *Domain*: used to identify the topic or subject of content.
- *Text analysis*: used to annotate content with lexical or conceptual information for the purpose of contextual disambiguation.
- *Locale*: specifies that a node is only applicable to certain locales.
- *Provenance*: used to communicate the identity of agents that have been involved in the translation of the content or the revision of the translated content.
- *External Resource*: indicates that a node represents or references potentially translatable data in a resource outside the document.

- *Target Pointer*: used to associate the node of a given source content (i.e., the content to be translated) and the node of its corresponding target content (i.e., the source content translated into a given target language).
- *ID Value*: indicates a value that can be used as unique identifier for a given part of the content.
- *Preserve Space*: indicates how whitespace is to be handled in content.
- *Localization Quality Issue*: used to express information related to localization quality assessment tasks.
- *Localization Quality Rating*: used to express an overall measurement of the localization quality of a document or an item in a document.
- *MT Confidence*: used to communicate the confidence score from a machine translation engine for the accuracy of a translation it has provided.
- *Allowed Characters*: used to specify the characters that are permitted in a given piece of content.
- *Storage Size*: used to specify the maximum storage size of a given content.

While these data categories describe a wide range of problem domains relevant when dealing with multilingual content, we focus on an application of named entity extraction and disambiguation to the *Text Analysis* data category.

The purpose of this is to annotate the results of text analysis localizable content. Besides translating names of entities, there are several other translation-related tasks and interactions that could benefit from NLP, for example:

1. Term suggestion: generating term candidates for terminology management.
2. Contextualization: providing context to localizable content that could get lost in the process.
3. Suggestion of things not to translate: for example, a specific named entity class should not be translated at all.
4. Automated transliteration of proper names: with certain language pairs, terms and proper names may be subject to special translation or transliteration rules.

In the requirements gathering phase of the standardization process, we had outlined a specification and a use case for the role of text processing components in the content localization process.

In general, the purpose of automatic annotation reduces the manual cost of annotation. For example, the enrichment of source content with named entity annotations is one example of such an automatic process. We define and standardize an interface for text processing components in a localization workflow, supported by a reference implementation.

The benefit of using text processing tools hinges on the adoption of a standardized interface that can lower the barrier to such a system. While named entity extraction has already been shown to improve translation quality [97], named entity extraction components are typically language-dependent. This often entails that they have different implementations, which increases the integration effort. The benefit of standardizing this interface is the reduced integration effort that needs to be applied for supporting text processing components for one additional language. Therefore, the standard defines a minimal-surface abstraction interface that supports the use case of text annotation in translation.

#### 3.4.6.1 Data and interaction model

In order to support the data model for these use cases, the ITS2.0 standard [122] introduces the *Text Analysis* data category. The Text Analysis data category is used to annotate content with lexical or conceptual information for the purpose of contextual disambiguation. This information is intended to be provided by text analysis software agents, such as named entity recognizers, lexical concept disambiguators, etc., and is represented by either string valued or Internationalized Resource Identifier (IRI) references to possible resource descriptions. For example: A named entity recognizer provides the information that the string "Dublin" in a certain context denotes a town in Ireland. While text analysis can be done manually, this data category is targeted more at software agents.

The information has several purposes, including, but not limited to:

- Informing a human agent such as a translator that a certain fragment of textual content (so-called "text analysis target") may follow specific translation rules. Examples: proper names, brands, or officially regulated expressions.
- Informing a software agent such as a content management system about the conceptual of textual entity to enable special processing. Examples: places, personal names, product names, or geographic names, chemical compounds, and protein names that are situated in a specific index.

Text Analysis support is achieved by associating a fragment of text with an external resource that a language review agent can interpret. The agent may for example use the web resource to disambiguate the meaning or lexical choice of the fragment, and thereby contribute to its correct translation. The web resource may as well provide information on appropriate synonyms and example usage. This is for example the case if the web resource is WordNet. In the case of a concept class, the external resource may provide a formalized conceptual definition arranged in a hierarchical framework of related concepts. In the case of a named entity, the external resource may provide a full-fledged description of the associated real world entity.

Extended example: we can enhance the word 'City' in the fragment 'I am going to the City' by one of the following:

- one of WordNet's [123] synsets that can be represented by 'city';
- an ontological concept of 'City' that could represent a subclass of 'Populated Place' as a concept;
- the central area of a particular city – as interpreted as an entity instance (e.g., 'City of London');

We can only annotate a given document fragment once. When support for multiple annotations is necessary (e.g., when we need to accomodate all three of the annotations in the extended example above) NIF 2.0 [124], TEI Stand-off Markup [125], or other so-called stand-off annotation mechanisms is better suitable. Some external resources such as DBpedia [110] also provide information for some ontological concepts and named entity definitions in multiple languages, meaning that named entity translation gets reduced to a knowledge base lookup.

Within the text analysis data category, the *its-ta-ident-ref* (identifier reference) property has been defined to represent any type of linkage between the annotated phrase and a knowledge base, making no assumptions about the type of the link.

#### 3.4.6.2 Example

The ITS2.0 standard allows annotating text using the following properties of the text analysis data category.

1. Text analysis confidence: The confidence of the agent (that produced the annotation) in its own result
2. Entity type / concept class: The type of entity, or concept class of the text analysis target IRI
3. Entity / concept identifier: A unique identifier for the text analysis target

Their usage in HTML5 is demonstrated in the following fragment:

```
<!DOCTYPE html>
<div
  its-annotators-ref="text-analysis|http://enrycher.ijs.si/mlw/toolinfo.xml#enrycher">
  <span its-ta-ident-ref="http://dbpedia.org/resource/Dublin"
        its-ta-class-ref="http://schema.org/Place">
    Dublin
  </span>
    is the
  <span its-ta-ident-ref="wordnet30:synset-capital-noun-3.rdf">
    capital
  </span>
    of
  <span its-ta-ident-ref="http://dbpedia.org/resource/Ireland"
        its-ta-class-ref="http://schema.org/Place">
    Ireland
  </span>.
</div>
```

Here, the *its-annotators-ref* attribute is used as a mechanism for annotating the provenance of certain ITS2.0 attributes, *its-ta-class-ref* is used to denote the type class of the entity behind the name 'Dublin', whereas *its-ta-ident-ref* is used to point to the DBpedia concept for the interpretation of the name Dublin. At the same time, the example also demonstrates the use of *its-ta-ident-ref* to point to an interpretation of the word 'capital'.

#### 3.4.6.3 Discussion

The reference implementation provided for standardization purposes uses the implementation of the named entity extraction and disambiguation techniques, presented in Section 3.4. These links to entities indicate whether a particular fragment of text represents a term, whether it is of a particular type, and alternative terms for that concept in other languages.

Best practices demonstrate that for a general purpose implementation, DBpedia can serve as a good knowledge base for entity identifiers, especially due to its multilingual nature. For word sense disambiguation, there are sense networks available for most languages, such as WordNet [123].

For domain classification, we use a general-domain taxonomy: the Open Directory Project [3] ontology. Classification itself is done by a large-scale hierarchical classification

[3][http://www.dmoz.org](http://www.dmoz.org)

approach [126]. The implementation of the ITS 2.0 processing code used in this subsection is available as an open source project [4].

Future work in this direction is in automating the annotation of other data categories, all of which are relevant for improving access to multilingual content. For instance, supporting the terminology lifecycle of extraction, tagging and mapping [127], or providing additional language and locale information of the content.

[4][https://github.com/tadejs/enrycher-its20](https://github.com/tadejs/enrycher-its20)

# Chapter 4

# Language in Social Media

Natural language processing has long relied on the correct grammatical structure of the language in order to make sense of it, as well as the fact that words have well-defined and known meanings. The emergence of new forms of social media in the last decade has given many people the means and opportunities to express themselves. However, language in social media can be grammatically very diverse. Certain types of social media, such as microblogging services encourage denser writing due to their limits, such as the 280-character limit on Twitter. Because of this limit, the content can be very difficult to understand since there is little space for the context, which is necessary to resolve ambiguities. For instance, methods for named entity disambiguation, presented in Section 3.4, cannot rely on heuristics that assume that every document is a longer piece of text that mentions multiple named entities, as well as interaction metadata. All these interactions give us a signature of how people interact with content, as well as making it possible to understand the content better. As a consequence, it can even help us understand the users' behavior better.

While in typical news and encyclopedic writing, the context scope is defined by the *document*, context in social media is hidden deeper: sometimes, it is known only to certain closely-related individuals, friends or insiders. In other examples, it can be inferred by looking at the author's previous posts, comments, and profile. On certain scenarios, such as sharing a link to a web page, it can be inferred from the content of the page they are sharing.

Social media content in various forms, such as self-expression, opinions on product or responses to real-world events have increasingly gained importance as they can provide a variety of different perspectives to the internet user. For instance, it can enhance a consumer's news reading experience, promote information sharing and even aid journalists in assessing their readership's response to a story.

In the case of integrating the readers' response to a story, the number of responses to an online news article may be huge. However, even in all the quantity of posts, there are not many points of view being presented, which opens a window of opportunity: if we see just a couple of well-chosen examples, we have covered the majority of the information value that the entire post stream contains on that topic. This presents us with a challenge of selecting only the most interesting responses for display. In Section of this chapter, we discuss an approach that poses this as a problem of sampling. We propose novel approach that outperforms known approaches from the literature on two evaluation set-ups, as well as an interpretation of the influence of various features.

Another property of text in social media that differs from typical corpora, such as news articles or encyclopedias, is that subjectivity and self-expression play a more prominent part. Naturally, this presents an opportunity for extracting that aspect of the content, so

that we may treat the subjectivity or polarity of the content dimension over which we can analyze our data, a task which is getting a lot of attention in the customer relation management (CRM) systems area. They are already adapting to the reality that the customers communicate with them through social media. Typically, here is where sentiment analysis systems come into play. While there has been significant work on sentiment analysis as a task by itself, our challenge is to develop a system that caters to the diversity of the input data that CRM systems need to deal with: in multiple languages, as well as multiple domains. While transfer learning approaches show some promise using domain and language adaptation, getting high performance that is typically required in such scenarios requires custom engineering. Since this re-engineering can be more expensive than data collection, our goal was to design a system that can be quickly adapted to novel domains, social media genres and languages with minimal fine-tuning.

In Section , we address the problem of sentiment analysis in an informal setting in multiple domains and in two languages and the implementation of such a system. Our main hypothesis is that if we model a given domain with multiple modalities, we get a more robust model that requires little re-engineering outside of acquiring training data.

We explore the influence of using background knowledge in the form of different sentiment lexicons, as well as the influence of various lexical surface features, described in . We evaluate several feature set combination strategies. We show that the improvement resulting from using a two-layer meta-model over the bag-of-words, sentiment lexicons and surface features is most notable on social media datasets in both English and Spanish. For English, we are also able to demonstrate improvement on the news domain using sentiment lexicons as well as a large improvement on the social media domain. We also demonstrate that domain-specific lexicons bring comparable performance to general-purpose lexicons.

## 4.1 Automatic Selection of Interesting Social Media Responses to News

Social media responses to news have increasingly gained in importance as they can enhance a consumer's news reading experience, promote information sharing and aid journalists in assessing their readership's response to a story. While many events happen and originate in social networks, our problem statement looks at the reverse perspective by observing responses to news articles describing real-world events. Given that the number of responses to an online news article may be huge, one challenge is to select only the most interesting responses for display. This section addresses this challenge by casting message selection as an optimization problem. We define an objective function that jointly models the utility scores of the messages and their entropy. Our solution first learns the utility of individual messages in isolation and then produces a diverse selection of interesting messages by maximizing the defined objective function. The intuition behind our work is that an interesting selection of messages contains diverse, informative, opinionated and popular messages referring to the news article, written mostly by users that have authority on the topic. Our intuition is embodied by a rich set of content, social and user features capturing the aforementioned aspects. We evaluate our approach through both human and automatic experiments in which we demonstrate that it outperforms the state of the art. Additionally, we perform an in-depth analysis of the annotated "interesting" responses, shedding light on the subjectivity around the selection process and the perception of interestingness. We analyze the different models and provide explanations on what are the important features on the level of individual messages. On the level of message sets, we employ a supervised pairwise ranking model to further demonstrate that individual message scoring alone does not bring out the best, but can be improved by diversification.

The work presented in this section is an extension of the work published in [128].

This section presents the following three contributions. First, it presents an analysis of the different scoring models and provides explanations for the effects the features have on the various indicators. Second, it presents and interprets a supervised model for predicting message set preferences, confirming that looking at message features alone does not provide optimal performance when selecting entire sets. Third, we report on an additional experiment that uses natural language processing techniques for textual entailment in place of a diversification model.

Overall, results presented in the literature point to a strategy that contextualizes the posted messages as much as possible, generating not just content, social and communication aspects, but also diverse textual and linguistic descriptors. To construct rich individual and collective features we therefore leverage insights from several existing approaches. Primarily, we build on existing work on sentiment and intensity analysis [58], [129] and apply basic redundancy detection [130]. We furthermore incorporate user authority, both for specific topics, as well as for topics in general [131]. We additionally explore the notion of diversity, which is also an important factor in how people perceive content [132] - whether diversity is important to a reader depends on on three main factors: the reader's prior convictions, emotional state, and social context.

Considering the large volume of messages posted in the context of news, keeping track of messages that refer to the most popular articles can easily become overwhelming. This information overload motivates the development of automatic systems that select and display only the most interesting messages. This *social media message selection* problem is illustrated in Figure 4.1: the selection system takes as input a news article and all the social media responses referring to that article, and outputs the most interesting subset of responses. Our work is motivated by a variety of existing and future applications in which

interesting social media responses could be automatically coupled with traditional news content, e.g. for displaying social responses near a news article for an enhanced reading experience.

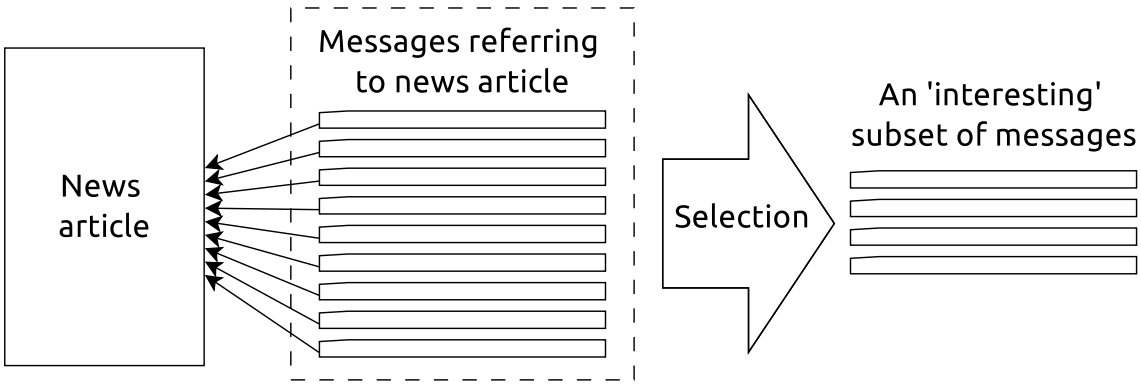


Figure 4.1: The social media message selection problem: selecting a small subset of messages sent in response to a news article.

Even though quantifying the interestingness of a selection of messages is inherently subjective, we postulate that an interesting response set consists of a diverse set of informative, opinionated and popular messages written to a large extent by authoritative users. By definition, interestingness of an object is the attitude characterized by a need to give attention to that object. While this is not something that we can directly measure in our data collection model, we can learn to estimate it based on the content itself and some observable interactions. By decomposing the notion of interestingness into these *indicators* we can pose the task of finding an interesting selection of messages as an optimization problem. We aim at maximizing an objective function that explicitly models the relationship among the indicators, thus producing selections that the typical person finds most interesting. We expect each indicator to have some influence on the messages ultimately selected to be part of the interesting set. Our method considers multiple content, social and user features to infer the intrinsic level of *informativeness, opinionatedness, popularity and authority* of each message, while simultaneously ensuring the inclusion of *diverse* messages in the final set.

We evaluate our approach through both human and automatic experiments, where we demonstrate that it outperforms the state of the art. In addition, we perform an in-depth analysis of the human evaluations, shedding light on the subjectivity and perception of interestingness in this particular task. To the best of our knowledge we have created the largest manually annotated news-response dataset to date, consisting of 45 news articles and 28,055 ratings per indicator annotated by 14 professional editors. Even though in this section we focus on messages posted on Twitter, our method can be applied to responses to news in other platforms, for instance article comments or discussion forums.

The remainder of this section is organized as follows. In Subsection 4.1.1 we present our proposed method, after which in Subsection 4.1.2 we describe our experiments and discuss the results. In Section 4.1.3 we then investigate the influences of the features on the various indicators.

### 4.1.1 Proposed method

We formulate the social message selection problem as follows.

**ALGORITHM 4.1:** Greedy iterative sampling algorithm

**Input** : A collection of messages $M$, the sample size $k$
**Output**: A sample set of messages $S^*$
$S \longleftarrow \{\arg\max_{m_i \in M} r(m_i)\}$
**while** $|S| < k$ **do**
  $m^* \longleftarrow \arg\max_{m_i \in M \setminus S} g(S \cup \{m_i\})$
  $S \longleftarrow S \cup \{m^*\}$
**end**
$S^* \longleftarrow S$

**Problem statement:** Given a news article and a set of related messages $M$, we seek a subset $S \subseteq M$ consisting of $k$ messages that are the most "interesting" to a typical reader in the context of the article.

We represent "interestingness" using a set of *indicators*. We identify four message-level indicators: *informativeness*, *opinionatedness*, *popularity*, and *authority*, as well as one set-level indicator: *diversity*. The intuition is that an interesting selection of messages contains *informative*, *opinionated* and *popular* messages referring to the news article. Furthermore, the messages should be written mostly by users who have *authority* on the topic. Looking beyond single messages, we posit that an interesting selection should contain messages that are *diverse* in content.

**Solution:** We computationally model the four message-level indicators using a utility function $r$ and the set-level diversity indicator using a normalized entropy function $H_0$. The solution to the social message selection problem is then to find the subset $S^*$ that maximizes the objective function $g(S)$, which measures the interestingness of subset $S$ as follows:

$$g(S) = \lambda \sum_{m \in S} r(m) + (1 - \lambda) H_0(S) \tag{4.1}$$

where $r(m)$ represents the utility score of a message $m$ and $H_0(S)$ is the normalized joint entropy of the entire set of messages $S$. We balance the effect of the collective diversity indicator on the sampling by specifying a suitable $\lambda$, which we will define manually in our experiments. The maximization problem is thus defined as follows:

$$S^* = \arg\max_{S \in 2^M, |S|=k} g(S) \tag{4.2}$$

In Section 4.1.1.2 we will describe how we derive $r$ and $H_0$, while we first focus on the sampling algorithm in Section 4.1.1.1.

#### 4.1.1.1 Sampling algorithm

We describe the sampling algorithm in Algorithm 4 and visualize it in Figure 4.2. The algorithm initializes by picking the element $m_i$ that ranks highest with the scoring function $r(m_i)$. Then, it iteratively adds the element $m_i$ that would provide the biggest boost in the objective function if added to the set $S$, stopping at the desired sample size $k$.

#### 4.1.1.2 Message scoring and message set selection

The two main components of the target function, $r$ and $H_0$, are both computed from a large set of features. These features serve as computational proxies for the five indicators

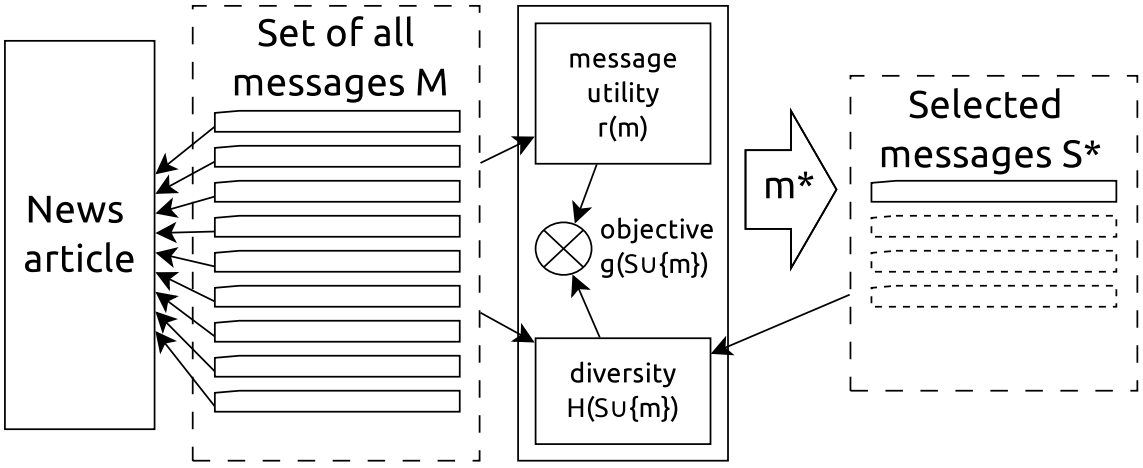


Figure 4.2: The diagram of the proposed algorithm, showing the two criteria, joint individual message and set diversity.

mentioned previously. Many of them are likely to be specific to the target social message platform; since in this paper we focus on Twitter, we therefore compute features specifically tailored to this platform, although most of them are nonetheless general.

**Features:** Table 4.1 reports the list of features we computed, where superscripts $s$ and $d$ indicate features that are used for computing $r$ and $H_0$, respectively.

- The *content features* capture the linguistic aspects that indicate how interesting, informative and opinionated a message is.
- The *social features* represent the level of diffusion of a message, reflecting its popularity.
- The *user features* capture the overall and topical authority of a user in the social network.

**Individual message scoring $r(m)$:** Given a tweet and its corresponding *scoring* feature values (scaled to interval $[0, 1]$), we use a supervised model as the scoring function $r(m)$. Specifically, we use the prediction given by support vector regression ($\epsilon$-SVR [133]), trained using the manually obtained labels for *interestingness*, as described in detail in Section 4.1.2.2.

**Entropy of message set $H_0$:** Given a message set $S$ and a set of *selection* features used to collectively select messages, we treat these features as binary random variables. Let $H_0(S)$ denote the normalized entropy of the message set $S$ given a probability model that uses the empirical probability of the selection feature values: $H_0(S) = H(S)/\log d$ in which $d$ is the number of binary random variables, and $H(S) = -\sum_{i=0}^{d} p(f_i = 1) \log p(f_i = 1)$, where $p(f_i = 1)$ is the empirical probability that the feature $f_i$ has the value of 1 given all examples in $S$. The intuition behind maximizing entropy is that it favors adding examples to $S$ with different non-zero features from the ones already in $S$, since adding an example with different features increases entropy more than adding an example already similar to the ones in $S$.

**Balancing individual and collective scores:** Using the parameter $\lambda$, we can balance the influence between scoring and entropy: $\lambda = 0$ yields an entropy-only approach, which we denote as ENTROPY, whereas $\lambda = 1$ gives a scoring-only approach that we refer to as SVR. The balanced combination at $\lambda = 0.5$ is denoted as SVR_ENTROPY.

**Using textual entailment:** Textual entailment is the property of a pair of statements, namely a hypothesis $H$ and a novel statement $T$, where $H$ entails $T$ if the truth value of $T$ follows from $H$. Here, textual entailment is used as a proxy for pairwise redundancy. To obtain the truth value of $entails(H,T)$, we use the EDITS textual entailment suite [134] with the best performing entailment model from the RTE6 competition [135], operating on an edit-distance model. In our scenario, we implement the entailment filter as a boolean function of a message $m$ and a currently selected set of messages $S$: if any of the currently selected messages in the set $S$ entails $m$, we do not select it:

$$\text{entails}(S,m) = \exists s \in S : \text{entails}(s,m) \tag{4.3}$$

The sampling expression then becomes:

$$m^* \longleftarrow \underset{m_i \in M \setminus S, \neg\text{entails}(S,m_i)}{\arg\max} g(S \cup \{m_i\}) \tag{4.4}$$

We denote the experiments using textual entailment with scoring as SVR_ENTAIL, and the combination of scoring, entropy and entailment as SVR_ENTROPY_ENTAIL.

#### 4.1.1.3 Comparison with the state of the art

Our research is closely related to the diversity-based sampling method presented in [48] and the social context summarization method presented in [64]. We first briefly describe these two methods and discuss their strengths and weaknesses in order to motivate our approach. From our review of the literature we observed that existing message selection approaches typically have a property to optimize for the diversity of the selection and/or for the utility of individual messages. In contrast with both these methods, we focus on a comprehensive notion of interestingness: whereas the diverse-based sampling method solely optimizes for tweet diversity, the social context summarization method focuses on optimizing for tweet informativeness. We therefore compare our proposed method against them in the evaluations.

Closely related to our problem is also the task of update summarization, which aims to create a summary of new information given one or more documents that have already been read, focusing mainly on web documents [140] and social media, especially on the topic of summarizing opinions [141]. An especially interesting problem is microblog conversation summarization, where the task is to make a summary of an entire tree of messages [142]; the authors demonstrated in this work that, besides content features, user influence information is also important for producing good summaries, demonstrating that messages from influential users should be more likely to appear in the summary.

*Diversity-based sampling*
The work in [48] tackles tweet selection using a greedy approach that iteratively picks tweets that minimize the distortion of entropy of the current subset, given a certain diversity criterion. The approach prescribes building a probabilistic model of tweets given their diversity using social, content and network features. The rationale behind the approach is that samples with a low entropy will be highly homogeneous with respect to the feature space, while those with a high entropy will be highly heterogeneous. In a social media message selection context it is thus preferred for the method to produce a sample with high entropy in order to promote diversity in the selected messages. The sample generation is initiated by adding a random message to the sample, after which new samples are iteratively added, so that the normalized entropy of the sample is closest to the pre-specified diversity parameter. This iterative process is repeated until the desired sample size has been reached. In our evaluations we refer to this method as DIVERSITY.

Table 4.1: The features used by our methods. Superscripts $s$ and $d$ indicate scoring (used in $r(m)$) and diversity (used in $H_0(S)$) features.

| **Content features** | |
|---|---|
| *N-grams*$^d$ | All unigrams, bigrams, and trigrams taken over the words in the tweet. |
| *Tf-idf score*$^s$ | Average tf-idf score of all words in the tweet, emphasizing rarely-used and penalizing out-of-vocabulary words. |
| *Log-likelihood*$^s$ | Likelihood of the tweet, based on a bigram language model constructed from all of the tweets of the article. |
| *Number of words*$^s$ | A higher number of words may indicate a tweet with more useful content. |
| *Me-information*$^s$ | Presence or absence of a first-person pronoun, indicating if the tweet is mainly about the author himself [136]. |
| *Question*$^s$ | Presence of a question mark in the content. |
| *Quote sharing*$^s$ | Presence of a quotation from the news article in the tweet. |
| *Quality*$^s$ | An measure of the tweet quality, indicating how well-written the tweet is. We use a supervised approach using a lexicon of low-quality terms, an English dictionary, and the proportion of words, hashtags, capitalized characters, and repetitions. |
| *Sentiment*$^s$ | % of positive and negative words, % of subjective words, and a mixed sentiment score, reflecting the presence of highly emotional or opinionated expressions in a tweet [129]. |
| *Intensity*$^s$ | A measure indicating the strength of the user's reaction [58]. |
| *Explicit controversy*$^s$ | A measure reflecting the mention of common controversial issues [129]. |
| *Location*$^d$ | The geographic location derived from the tweet, either from a mentioned place or the location of the user. |
| **Social features** | |
| *Retweet and reply*$^{s,d}$ | A flag indicating whether the tweet is a retweet or a reply. |
| *Followers, friends*$^s$ | Number of followers and friends. |
| *Follower-friend ratio*$^s$ | Ratio between the number of followers and friends. |
| *Number of retweets*$^s$ | The total number of retweets of the user's posts over the time of data gathering, signaling user authority [137], [138]. |
| *Number of users retweeting*$^s$ | The total number of other users that retweeted this user at least once. |
| *Tweet-retweet ratio*$^s$ | Ratio between the number of tweets a user posts and the number of retweets received. |
| **User features** | |
| *User verified*$^s$ | A flag indicating if the user is verified by Twitter, which may increase the credibility of the user's posts. |
| *User spam*$^s$ | A flag indicating if the user is a spammer, based on a large spam dictionary trained on web page spam. |
| *User authority score*$^s$ | A global authority score, computed to have a topic-independent estimate of the user's overall importance. |
| *User topic authority*$^s$ | We consider a user $u$ authoritative for the article's topic, when previous tweets on the same topic were often retweeted. We extract the three most relevant named entities from the article using our "aboutness" system [139] and consider these entities as a synthetic summary of the topic of the article. For each of these entities we then extract the set of tweets of the user that mention the entity and look up the number of times any tweet in this set was retweeted by other users. The topic authority score is then computed as a combination of the relevance of the entities for the article and the number of retweets. |

*Social context summarization*
The work in [64] proposes an approach based on conditional random fields. It simultaneously addresses message and document summarization by modeling messages and sentences as connected 'wings' in a factor graph, where factors are assigned to individual message and sentence features, between their instances on a single wing and across both wings. To determine the key sentences and the important messages, an approximate inference approach iteratively updates and propagates the beliefs between factors and labels, where the final labeling indicates which sentences and messages are selected. We refer to this dual-wing factor graph method as DWFG. In the same paper, a simplified model was also presented, implemented as a logistic regression classifier, that enriches each tweet with the features of the sentence most similar to it, rather than introducing sentence-tweet factors between similar tweet-sentence pairs as is the case in the original model. In the task of summarizing relevant tweets for a news article this method yielded good performance. We refer to this method as LR+.

*Observations*
Even though the entropy measure used in the DIVERSITY approach has been demonstrated to be a good proxy for diversity, it does not evaluate individual messages with respect to their individual characteristics. Furthermore, its content probability model assigns each message to a broad category, such as 'politics' and 'sports'; while this works very well for obtaining a diverse sample on broader topics, this may not be a good fit in the case of responses to a single news story on a relatively narrow topic. The dual-wing factor graph and logistic regression methods take a richer set of features into account, but do not explicitly model diversity in the context of how people respond to news. Specifically, connecting messages only through their reply/retweet relations means that highly similar or even duplicate messages, which are a common occurrence in the context of news, may not be linked to each other, likely resulting in a substantial amount of redundancy being present in the final selection. Our method therefore offers the following improvements:

- We focus on a comprehensive notion of interestingness instead of solely optimizing for message diversity (DIVERSITY) or for message summarization (DWFG and LR+).
- We detect and avoid redundancy between the content of messages by including a richer set of textual features.
- We model the importance of messages beyond their literal content by including features that capture the wider social, conversational and linguistic context, e.g. the amount of opinion, emotion and controversy of a message, the level of authority of the author on the topic, and the user's intent of the message [136].

### 4.1.2 Experiments

In this section we describe our experiments for evaluating the performance of our proposed method among variations of itself, and against the previously mentioned baseline approaches on the problem of selecting the most interesting tweets posted in response to news.

#### 4.1.2.1 Dataset

We collected a set of tweets posted on Twitter between February 22, 2011 and May 31, 2011 that were written in the English language and that included a URL to an article published online by news agencies such as Reuters, Huffington Post, India Times and Al Jazeera. We crawled each of these links to obtain the original news article content, discarding

redirected and unresolved links. We observed that the news-related tweets frequently had highly similar content to each other; when ignoring retweet markers, links, mentions and hashtags, and only purely looking at the words, just over half of the tweets were identical in our dataset. For each news article, we therefore filtered out all such duplicate tweets and only retained the one with the earliest timestamp. For the experiments we ultimately picked a set of 45 news articles that spanned a wide range of topics, including business, politics, technology, weather, disasters, and warfare. We ensured each article had on the order of 100 unique tweets associated with it, so that there would not only be sufficient tweets for the evaluated methods to work with, but also so that the tasks would not overly exhaust the annotators that created the gold standard.

While the minimum unique message requirement is necessary for collecting sufficient amounts of data for the experiment, it is biased towards articles that may elicit more engagement from the viewers than they might have otherwise.

#### 4.1.2.2 Gold standard collection

We asked a team of 14 annotators to create a gold standard collection by: a) rating the individual tweets sent in response to an article on the *informative* and *opinionated* indicators, and b) selecting approximately 10 tweets they found the most *interesting* as a set in the context of the news article. We did not ask the annotators to rate tweets for the *authority* and *popularity* indicators due to the difficulty of assessing these values at the tweet level and without context. The annotators were Yahoo! News editors familiar with the topics relevant to their respective areas. Despite their specializations, all of the annotators were asked to annotate tweets across all news articles. Instead, we opted for a simple mechanical model for deriving the values of these indicators, as we will explain later in this section.

For the *informative* and *opinionated* indicators, the annotators were asked to assign a score to each tweet on a scale from *1* (the tweet decidedly does not exhibit the indicator) through *2* (the tweet somewhat exhibits the indicator) to *3* (the tweet decidedly exhibits the indicator), while allowing them to skip tweets that they considered to be either spam or not relevant to the article. To strengthen the gold standard collection, each task was assigned to three different annotators. Due to time constraints, the number of ratings performed by annotators varied. In total, 28,055 ratings were gathered, of which approximately 70% were *1*s, 15% were *2*s, and 15% were *3*s. In other words, in aggregate, 70% of the tweets were not considered informative nor opinionated by the annotators.

As previously mentioned, we did not ask annotators to rate tweets for the *authority* and *popularity* indicators. First, the authority indicator is not straightforward to assess without context: while identifying major news outlets as authoritative (e.g., AP, Reuters, BBC, etc.) might be trivial, doing so for individuals is more nuanced. Second, the popularity indicator is directly related to how often a tweet is retweeted and replied to, which is difficult to assess when seeing a tweet by itself. Therefore, we opted for a simple mechanical model for both: for the authority indicator we combined the values of user authority and topic authority features of the users writing the tweets, and for the popularity indicator we combined the retweet and reply counts for each tweet.

*Inter-annotator agreement* We measured the inter-annotator agreement for all three tasks using Cohen's linearly weighted kappa statistic [143]. We averaged the pairwise kappa values of all possible combinations of annotators that had overlapping tweets in those they had rated, obtaining the overall kappa. Because the annotators had different scoring behaviors, i.e. they used different amounts of *1*s, *2*s and *3*s from each other, the maximum attainable kappa was less than 1.0. To place the kappa into context, we normalized it by dividing the obtained kappa by the maximum obtainable kappa.

Even though the definition of what constitutes a 'good' kappa varies across the literature, a higher kappa always indicates higher agreement. Overall, the annotators had a relatively high agreement on *opinionated* tweets with $\kappa_{cwopi} = 0.61$ (with maximum obtainable value of 0.70), while only fair agreement of $\kappa_{cwint} = 0.35$ (with maximum obtainable value 0.57) for *interesting* tweets and just slight agreement of $\kappa_{cwinf} = 0.20$ (with maximum obtainable value of 0.56) for *informative* tweets. The agreement on the latter two indicators indicates a high subjectivity. The annotators approached the content with different background knowledge, which may have contributed to what they themselves considered informative. As an example of inter-annotator (dis)agreement, we show three tweets sent in response to an article about Japanese citizens returning cash they discovered in the rubble after their town was hit by a tsunami:

> *"I am continually amazed at the honor and trustworthiness of the Japanese people. Truly astounding and heartening to see!"*
>
> *"Japanese citizens turning in cash found in tsunami zone - #cnn"*
>
> *"Giving respect to the Japanese people."*

The annotators agreed that the first tweet was interesting, possibly due to the personal touch of this well-written tweet. They also agreed that the second tweet was not interesting, presumably because it repeated the title of the news article. However, the annotators disagreed on the last tweet, where one annotator may have agreed with the conveyed idea, whereas another may have considered it someone's opinion and therefore not of interest. The latter case exemplifies that rating tweets is a subjective matter.

#### 4.1.2.3 Experimental evaluation

Using the gold standard annotations, we evaluate the methods on predicting interestingness. We also evaluate the performance on opinionatedness and informativeness in order to estimate their difficulty.

**Methods:** We evaluated the baseline methods DIVERSITY, DWFG and LR+ as presented in Section 4.1.1.3 against variations of our proposed method: SVR, ENTROPY, SVR_ENTROPY, SVR_ENTAIL, and SVR_ENTROPY_ENTAIL.

We also experimented with other scoring mechanisms for individual messages, such as using the likelihood estimates from a logistic regression classifier for probabilistic scoring, and with other diversity mechanisms at the collective level, such as using more sophisticated natural language processing methods, like group-wise textual entailment. However, these alternatives did not show any significant benefit, and we therefore do not include them in our evaluations.

**Training:** The methods that required training were trained with negative and positive judgments. For interestingness positive examples are those included in the interesting set by the annotators for each article, and all others are negative. For opinionatedness and informativeness positives and negatives are respectively ratings *3* and *1* from all annotators on all articles. We randomly sampled the tweets to ensure a balanced distribution of positive and negative judgments. For the DWFG we labeled all sentences of the article as *positive*, since we can consider a news article to be a summary of a newsworthy event. The hyperparameters we tuned on a tuning set of 5 news articles and their messages. Where available, hyperparameters of baselines used the values as reported by the original authors.

**Evaluation measures:** To measure performance, we computed both the ROUGE-2 [144] and $F_1$ scores for the *informative*, *opinionated* and *interesting* indicators. These scores

measure how closely the output of a method is to the idealized output, i.e. those scored positively by the annotators. While $F_1$ only considers exact tweets to be correct, ROUGE-2 counts tweets with similar content as a partial match, reporting the recall of word bi-grams of the obtained sample and the gold standard examples. We use the ROUGE-2 variant because it correlates best with human judgment among all ROUGE-N variants [145]. We performed 10-fold cross-validation, using nine folds for training (when needed) and one for testing.

**Results:** Table 4.2 contains the summary of our experimental results. While our proposed methods outperform the DIVERSITY, DWFG and LR+ baselines on the main task of predicting interestingness, the two evaluation metrics show different rankings of the methods. Since the baseline methods are either unsupervised or do not explicitly train for all the criteria, its corresponding scores may be lower in those cases.

- For *interestingness*, the best ROUGE-2 performance was obtained by methods that have an entropy selection component (ENTROPY, SVR_ENTROPY), while $F_1$ favors most SVR-based approaches.
- For *opinionatedness*, we observed that while the ROUGE-2 scores were similar across ENTROPY-based methods, $F_1$ scores were again higher among most SVR-based approaches. Although the margin was not statistically significant, the LR+ baseline was able to achieve the highest ROUGE-2 performance, likely due to enriching the feature set with the features of the most similar sentence. However, even though ROUGE-2 measured that the lexical overlap with the correct messages was high, $F_1$ scores show that the actual correct messages were not selected as often.
- For *informativeness*, the overall results were lower than for other indicators. This result matched the lower inter-annotator agreement of the ratings. Upon inspection we noticed that for the most part, informative tweets did not contain article snippets, but instead contained novel text. We also noted the same pattern of ROUGE-2 favoring ENTROPY, while $F_1$ favoring SVR.

Table 4.2: Comparison of the tweet sets generated by the methods in terms of their informativeness, opinionatedness and interestingness as measured by ROUGE-2 and $F_1$ at $k = 10$. The asterisk marks the cases where the measurements were significantly higher than all of the three baselines, having a T-test p-value below 0.05.

| | Informative | | Opinionated | | Interesting | |
|---|---|---|---|---|---|---|
| | R-2 | $F_1$ | R-2 | $F_1$ | R-2 | $F_1$ |
| DIVERSITY | 0.178 | 0.050 | 0.270 | 0.108 | 0.280 | 0.109 |
| DWFG | 0.170 | 0.048 | 0.259 | 0.056 | 0.235 | 0.051 |
| LR+ | 0.175 | 0.032 | **0.355** | 0.132 | 0.264 | 0.054 |
| ENTROPY | **0.246*** | 0.124* | 0.345 | 0.184* | **0.357*** | 0.216* |
| SVR | 0.212* | **0.132*** | 0.326 | **0.217*** | 0.336* | **0.239*** |
| SVR_ENTROPY | 0.232* | 0.123* | 0.343 | 0.211* | 0.346* | 0.224* |
| SVR_ENTAIL | 0.218* | 0.120* | 0.311 | 0.210* | 0.339* | 0.228* |
| SVR_ENTROPY_ENTAIL | 0.222* | 0.115* | 0.338 | 0.184* | 0.339* | 0.193* |

To summarize, the ENTROPY approach excels at content fragment retrieval (ROUGE-2), whereas approaches combining a scoring and an entropy selection model are able to provide better performance on $F_1$, offering a tunable trade-off between diversity and individual tweet scores. The use of textual entailment, however, does not show any consistent

improvement over the use of entropy sampling as a diversification strategy. None of the three tasks show measurable benefit neither from incorporating textual entailment instead of entropy sampling, nor from combining them.

We furthermore noticed some key differences in behavior of the baselines we compared against. Whereas the DWFG method ensures content diversity by encoding the diffusion network into the factor graph, so that retweets do not get selected when the original tweet does, in the news domain many near-duplicates tweets were actually not retweets at all: when responding on the same article, several users posted very similar responses independently, resulting in lower performance on ROUGE-2. While this is not captured by the dual-wing factor graph design, it is implicitly taken care of in DIVERSITY, where it diversifies the tweet content represented with a set of topics. However, the original DIVERSITY scenario focused on sampling from tweet sets that are retrieved via a hashtag query, which are still diverse enough to be accurately described by topics. On the other hand, news responses tend to be topically narrower, which favors our content representation for calculating sample entropy: instead of probabilities of topic annotations, we use the probabilities of word n-grams, which allows a finer-grained model, since most of the responses to a news article tend to be on the same topic, but still have different messages.

*Understanding tweet "interestingness"* Given the gold standard annotations, we aimed to infer to what extent we can decompose the individual *interestingness* of a tweet into the individual indicators and how much influence each of the indicators had. In this investigation, we averaged the ratings received or computed for the individual indicators. We used the gold standard labels for the *informative*, *opinionated* and *interestingness* indicators as well as the automatically derived labels for the *popularity* and *authority* indicators in the same fashion as in Section 4.1.2.2. After normalizing indicators to the same interval, we fitted a least-squares linear model and obtained the following linear combination, supported by a coefficient of determination $R^2$ of 0.38, having a correlation coefficient of 0.62:

$$int = 0.60 \cdot inf + 0.29 \cdot opi + 0.03 \cdot pop + 0.10 \cdot aut \tag{4.5}$$

While making an assumption of independence of these factors, a linear model explains only 38% of the variance and it gives us a glimpse of what constitutes interestingness. The primary driver is *informativeness*: while our annotators had low agreement on how informative a tweet is, having at least one of them mark a tweet as such was sufficient for the signal to get picked up. The second driver was *opinionatedness* followed by *authority* and *popularity*. We can also explain the low coefficient for popularity by the fact that it strongly correlates with the authority indicator, leading to a single coefficient dominating in a linear model where variables are correlated.

#### 4.1.2.4 Preference judgment collection

While we have already shown a promising approach compared to existing literature, we conducted further experiments in order to better understand the differences among the variations of our proposed methods and to understand the extent to which certain factors contribute to the interestingness of a set of tweets. We performed a second evaluation in which we used another team of annotators, none of whom had previously participated in the gold standard collection, using the crowdsourcing service CrowdFlower, currently known as Figure Eight.

We evaluated the satisfaction of the annotators with the sets of tweets ultimately produced for the same collection of 45 news articles. The number is relatively low due to annotation constraints. Instead of asking the annotators to assign a numeric score to

each set of tweets, we paired the methods and placed them in a head-to-head competition. Annotators were then asked to express preference for either set or to indicate whether they were equivalently interesting. This approach allows us to extrinsically evaluate the quality of the sets produced by each of the methods, rather than having the annotators score each set separately or pick from multiple sets at once, both of which could have given ambiguous results. Unlike ratings, comparative judgments generate highly reliable responses and clear individual differences.

We requested at least five judgments for each task and appropriately mixed in trivial gold standard scenarios to filter out malicious annotators, where we selected only judgments from annotators who had at least 75% precision on the gold filter questions. The reason for the 75% gold filter threshold for annotator selection is due to past experiences with crowdsourced annotations, specifically the fact that a portion of the annotators don't always answer honestly, but merely choose randomly or pick the first available answer among several adversarial scenarios. While research in [146] describes several methods for identifying and filtering adversarial workers, their baseline is a 60% gold filter threshold. Furthermore, Crowdflower (Figure Eight) itself recommends setting the threshold to minimally 70% [147].

After expressing a preference, the annotators were presented with a follow-up question to indicate why they had chosen one set of tweets over the other, allowing us to compare their preferences with the ratings the previous annotators had given to see what kind of, if any, correlations exist between the individual tweet and collective selection indicators, and the interestingness of the samples.

#### 4.1.2.5 Preference judgment analysis

Once again, we first analyze the agreement between annotators and then look at the outcome of the evaluation to see which method performed better than the others. Afterwards, we take a closer look at what constitutes an interesting selection of tweets. We structure our analysis as follows:

**Combining diversity and scoring:** We are interested in understanding whether diversity alone or scoring alone is enough for achieving an interesting selection.

**Using textual entailment for diversity:** We are interested in understanding whether it is worth employing textual entailment instead of entropy to ensure non-redundancy of content.

**Comparison against baselines:** Using our best performing method so far, we are interested in knowing how it compares to the existing baselines on the same problem.

In order to limit the number of pairwise comparison evaluations when comparing against baselines, we chose to compare our best-performing method in the first two analyses against DIVERSITY and DWFG, the principal methods proposed by the authors in their respective papers.

*Inter-annotator agreement* We measured the inter-annotator agreement for the preference judgments in two ways. We first used Cohen's unweighted kappa statistic to calculate all pairwise kappas when there was some overlap in the tasks between two annotators, which we averaged to obtain the overall $\kappa_{cu}$ of 0.50, having a maximum $\kappa_{cumax}$ of 0.81 and thus yielding a mean normalized $\kappa_{cunorm}$ of 0.62, which shows moderate agreement. However, since every annotator only judged a limited number of examples and the overlaps were relatively sparse, we also measured a multi-rater Fleiss' kappa [148]. Even though the overall $\kappa_f$ was fair at 0.21, the results varied across different comparisons. While the

Cohen's and Fleiss' kappas are not directly comparable, both point to fair-to-moderate overall agreement. Agreement measurements in the final column of Table 4.3 show that in the cases with higher agreement, the results of the method comparison were also clearer (with some exceptions). For instance, ENTROPY vs. SVR_ENTROPY demonstrated low agreement, but still had a significant difference in overall votes, meaning that the ratio of votes was more constant over all articles, as opposed to there being a particular article where one method would consistently dominate. The cases with moderate agreement, such as SVR vs. SVR_ENTROPY and DIVERSITY vs. DWFG highlight situations where one method can objectively produce better samples.

Table 4.3: Comparison of tweet selection methods based on annotator preference votes, along with inter-annotator agreement via Fleiss' kappa $\kappa_f$. Scores from 0.00 to 0.40 indicate slight agreement, while 0.40 to 0.75 indicate good agreement.

| Method A | Method B | Votes A | Votes B | None | $p$-value | $\kappa_f$ |
|---|---|---|---|---|---|---|
| **Combining entropy with scoring** | | | | | | |
| SVR | SVR_ENTROPY | 151 | **194** | 16 | 0.02 | 0.35 |
| SVR | ENTROPY | **193** | 155 | 9 | 0.04 | 0.25 |
| ENTROPY | SVR_ENTROPY | 114 | **147** | 13 | 0.04 | 0.06 |
| **Using entailment for diversity** | | | | | | |
| SVR_ENTAIL | SVR_ENTROPY | 91 | **116** | 4 | 0.08 | 0.04 |
| **Comparison against baselines** | | | | | | |
| DIVERSITY | SVR_ENTROPY | 79 | **158** | 3 | 0.00 | 0.04 |
| DWFG | SVR_ENTROPY | 99 | **149** | 5 | 0.00 | 0.09 |
| DWFG | DIVERSITY | **118** | 94 | 10 | 0.10 | 0.48 |

#### 4.1.2.6 Method comparison

We aggregated the preference counts of the annotators, which are also shown in Table 4.3. We conducted a $\chi^2$ statistical test over the aggregate counts with the null hypothesis that both of the methods in a given pair are equivalent.

We summarize the results of the three experiments as follows:

**Combining diversity and scoring:** We observed that a combination of diversity and scoring (SVR_ENTROPY) produced better selections than diversity (ENTROPY) or scoring (SVR) alone. When evaluated separately, scoring (SVR) was better than diversity (ENTROPY), i.e. in our dataset, where the tweets were already relatively diverse due to the prefiltering of duplicate tweets, scoring proved to be more important.

**Using textual entailment for diversity:** When analyzing the performance of using textual entailment instead of entropy, we noticed a slightly lower preference for samples obtained by diversification via textual entailment, which is consistent with the results from the gold standard analysis. We noticed that textual entailment was difficult to apply to the informal language used in tweets, so the cases where it would have proven effective were outnumbered by the cases where entropy sampling was already robust enough. Since the difference was small, we speculate that it could be useful in combination with text corrections [149].

**Comparison against baselines:** Using our best performing method so far, we are interested in knowing how it compares to existing baselines on the same problem. Even though SVR_ENTROPY was consistently good, albeit not always the best within the gold standard evaluation, in the light of the preference judgment results we do consider it our preferred option for the comparisons against the baselines. Looking at the head-to-head comparisons against the baselines, we observed that the DIVERSITY approach performed similarly to the DWFG approach, whereas both were outperformed by our proposed method SVR_ENTROPY.

*Understanding set "interestingness"* After the annotators had completed a pairwise comparison task, we posed a follow-up question by asking the annotators what they considered the reason for their preference. They answered a multiple-choice question selecting the indicators they considered that contributed to interestingness. We omit popularity and authority from the selection since we do not expect users to infer them from the situation at hand, so both were calculated using the same process as in Section 4.1.2.2. If one sample had at least one more popular or authoritative tweet than the other, that indicator was automatically counted as a reason. While this does not enable comparison between importance of individual indicators, it does allow for comparisons across various method pairs. Users were also asked to supply additional reasons beyond the predefined ones.

Table 4.4 shows the number of times the annotators stated a particular reason for preferring the selection produced by one method over the other, ordered by the preferred method within that method pair. While no data exists on the confidence of their judgments outside of counting the number of cases no choice has been made, we can still interpret a high frequency for a given reason as a point of distinction among methods. For instance, annotators preferred SVR_ENTROPY over SVR mostly on the fact that it provided more opinionated samples, while the opposite case was more evenly distributed among different reasons.

The most interesting observation is the prominence of opinionatedness in some scenarios: this was the primary driver between distinguishing SVR and ENTROPY, demonstrating the effectiveness of having the scoring operate on sentiment and intensity features. On the other hand, informativeness was more difficult to predict, making the samples harder to distinguish on informativeness than on opinionatedness, where differences are more obvious. Authority was less prominent in the aggregate, but was often the primary driver when the annotators chose DIVERSITY over another method. In some cases, the annotators remarked that they knew the author, suggesting that the social proximity of the author of the tweet to the annotator (and subsequently any reader of the produced selections) serves as a potential personalization opportunity. Other comments given by the annotators outside of the predefined choices noted humor as an important factor for preferring one set over the other, which was not modeled by any of the methods, as well as several instances of an annotator picking a set due to the profanity or too much controversy in the other set. This suggests the importance of incorporating features that capture this.

While textual entailment has initially shown promise to be a basis of a better diversity mechanism, we evaluated the combination of textual entailment and entropy as an ensemble (SVR_ENTROPY_ENTAILMENT) in initial experiments and it did not perform sufficiently well to warrant its inclusion in the pairwise preference data collection.

### 4.1.3 Feature analysis

#### 4.1.3.1 Individual message features

In order to better understand what drives annotators to judge something as interesting, informative, or opinionated, we provide an interpretation of the trained scoring models.

Table 4.4: Distribution of preference reasons for the preferred method in a given method pair.

| Method A | Method B | Preferred | inf. | opi. | aut. | pop. |
|---|---|---|---|---|---|---|
| SVR | SVR_ENTROPY | SVR_ENTROPY | 80 | 115 | 31 | 33 |
| | | SVR | 74 | 78 | 28 | 16 |
| SVR | ENTROPY | SVR | 75 | 124 | 23 | 23 |
| | | ENTROPY | 62 | 98 | 20 | 31 |
| ENTROPY | SVR_ENTROPY | ENTROPY | 66 | 54 | 16 | 10 |
| | | SVR_ENTROPY | 88 | 69 | 8 | 3 |
| SVR_ENTAIL | SVR_ENTROPY | SVR_ENTROPY | 87 | 52 | 9 | 14 |
| | | SVR_ENTAIL | 65 | 49 | 17 | 6 |
| DIVERSITY | SVR_ENTROPY | SVR_ENTROPY | 100 | 76 | 8 | 3 |
| | | DIVERSITY | 47 | 46 | 58 | 65 |
| DWFG | SVR_ENTROPY | DWFG | 58 | 58 | 82 | 56 |
| | | SVR_ENTROPY | 88 | 76 | 3 | 12 |
| DWFG | DIVERSITY | DIVERSITY | 76 | 49 | 20 | 46 |
| | | DWFG | 88 | 59 | 46 | 4 |

We do this by performing feature selection that considers the contributions of individual features to the $F_1$ score of the scoring model. To achieve this, we apply recursive feature elimination using cross-validation [150], an approach that, given a linear model, removes the feature with the lowest weight at every iteration, while simultaneously validating the model performance.

We also consider the direction in which individual features are pulling: while some features contribute positively to the overall score, some have a negative impact. While observing the feature weights alone is only feasible in linear models, we can interpret non-linear models with a technique such as LIME [151] by estimating a locally linear model around a certain data point.

Figure 4.3 shows the various contributions of individual features. The horizontal axis represents the feature weight coefficient in the linear model: positive weights contribute to higher scores, while negative values pull the overall scores down. We denote features belonging to the best performing subset of the recursive feature elimination procedure in bold font. The feature elimination highlights which features are important in a way that measurably affects prediction performance. For example, if there are features that are correlated amongst themselves and carry the same signal, only one of these features would get selected, since adding the others would not further improve performance.

For interestingness, the log-likelihood given the language model has the strongest negative signal, suggesting that repetition of the same phrases compared to the overall message set is undesirable. An important positive feature is the relative proportion of words to non-word tokens, followed by the ratio of friends to followers and the sentiment score, biasing the selection towards positive sentiment. Important negative signals are also length, log-likelihood, the model-based quality estimate, and the poster's social standing.

Informativeness has similar positive features to interestingness, while the main differences are in the stronger focus on user authority, more textual content and, surprisingly, a moderate amount of controversy. Negative indicators include high log-probability given the entire message set, me-information, and junk word percentage.

The weights for opinionatedness show that there is more to expressing an opinion than just using terms with a sentiment value: the subjectivity indicators, linguistic quality and

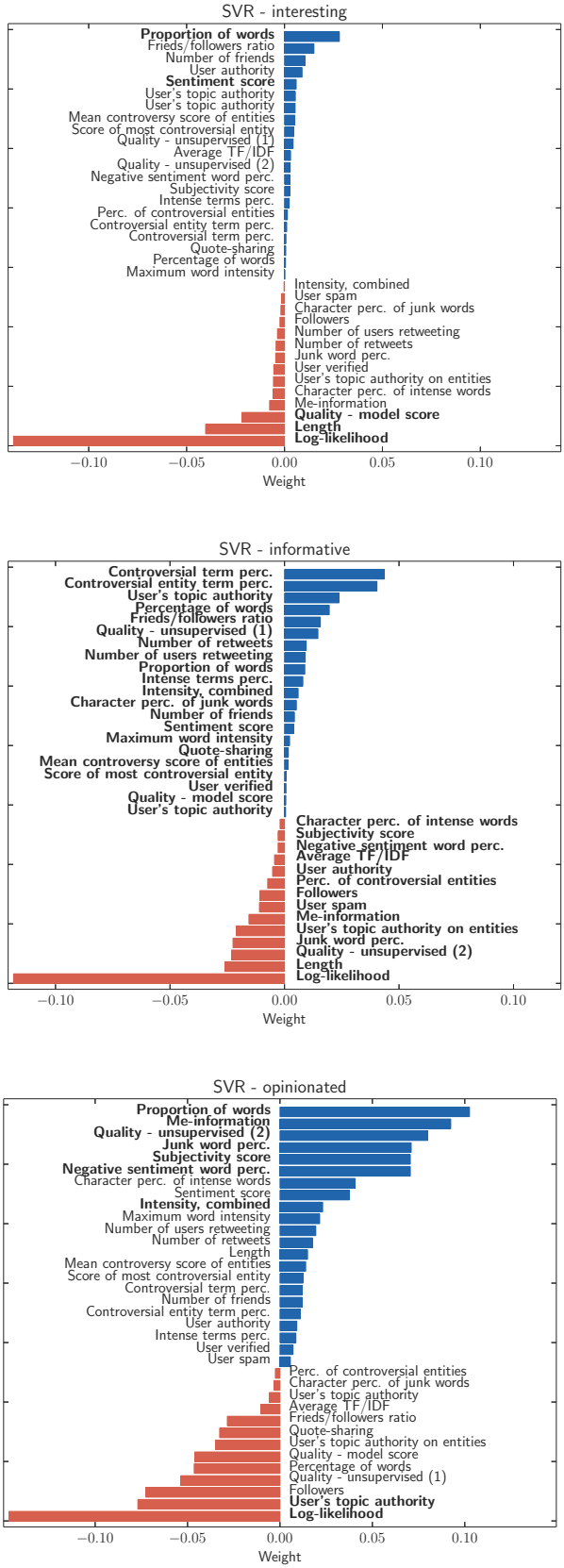


Figure 4.3: Importances of features as linear coefficients of the scoring model. Bolded names represent important features, chosen by recursive feature elimination.

me-information are also what one would intuitively expect to be present in an opinion, as well as the intensity of the expression. This example also demonstrates the differing influence of quality model features: while one is an important positive feature, another is a moderate negative feature, suggesting that only certain narrower definitions of quality are predictive of opinionatedness. On the other hand, controversy does not play a major part. The feature elimination suggests that we can detect opinionatedness with a smaller set of features: none of the controversy features have a significant impact on the prediction quality.

#### 4.1.3.2 Learning to rank message sets

Besides providing an analysis based on gold-standard annotations, we also look into the differences of the aggregate features of whole message sets, in order to show what makes one message set more likely to be chosen than another in a preference judgment scenario as described earlier in Section 4.1.2.4. Using the data gathered in the preference judgment validation, we aim to train a predictor that is able to pick the more interesting set.

To achieve this, we pose this as a learning-to-rank problem, where each of the 1,802 preference judgment pairs from our experiment is considered as a data point within the ranking problem: the chosen message set ranks first, while the other one gets second place. This setting is also known as ordinal regression, which has successfully been used in combination with Support Vector Machine classifiers to predict rankings [152], [153] and can be efficiently solved using the online Pegasos [154] approach via stochastic gradient descent [155]. The basic approach behind the ordinal ranking SVM is to take every pair of comparable elements, calculate their difference in the feature space and treat this difference as a data point, where the label 1 is assigned when the first element ranks higher and -1 otherwise. In line with our experiments presented in previous sections, we hypothesize that the pairwise ranking model should not just take an aggregate of individual message features into account, but also the diversity of the entire set. We evaluate the following feature construction protocols:

**Feature averaging ($\mu$):** For every message set, we compute the mean of the individual message features as its feature vector.

**Feature averaging, combined with standard deviations ($\mu$-$\sigma$):** In addition to the mean, we also compute the standard deviation of each feature in the sample, using that as a new feature of the entire sample.

**Feature averaging, combined with entropy ($\mu$-$H$):** We include the normalized entropy of the message set, as described in Section 4.1.1.2.

**Feature averaging, combined with standard deviations and entropy ($\mu$-$\sigma$-$H$):** We include all the aforementioned feature sets: mean of features, standard deviations of features, and sample entropy.

We compare these feature sets in a ranking pairwise ranking prediction setup. A random baseline would result in 50% accuracy. The results in Table 4.5 show that using either the variance or the entropy both improve the ranking performance by a small margin compared to using feature means alone. The margin is statistically significant against the averaging-only $\mu$ baseline in the cases of $\mu$-$\sigma$ and $\mu$-$\sigma$-$H$. It also opens up additional possibilities of learning not only selection features, but also diversification features by observing the features with high variance. Since the data used in this comparison also included samples

Table 4.5: Performance of pairwise ranking prediction on various feature sets. The bold text indicates that the result is significantly higher than the RankSVM-$\mu$ baseline.

| Approach and feature sets used | Accuracy |
|---|---|
| RankSVM-$\mu$ | 0.669 |
| RankSVM-$\mu$-$\sigma$ | **0.687** |
| RankSVM-$\mu$-$H$ | 0.672 |
| RankSVM-$\mu$-$\sigma$-$H$ | **0.676** |

generated by baselines, sample entropy does not necessarily carry any positive signal, resulting in mediocre performance of $\mu$-$H$.

Figure 4.4 shows the features of the combined RankSVM-$\mu$-$\sigma$-$H$ model that are distinctive for the ranking problem at hand. While features were excluded due to a large amount of noise in the dataset, the standard deviation aggregated features are indicated as predictive, especially in terms of length and quality model scores. Intensity was especially prominent as a positive feature, even more than in the individual tweet models. Interestingly, quote sharing and length are seen as negative features, possibly contributing to a stronger information overload when presented with a set of messages compared to a single-message setting.

Figure 4.4 also shows the important factors that influence pairwise preferences. However, these pairwise ranking models do not necessarily generalize well across different sampling methods, given the amount of training data we had available. For instance, across the entire dataset of evaluated method pairs, entropy is not universally a positive feature, as seen in Figure 4.4, but gets a positive weight when training only with sample pairs that were generated from approaches that included a scoring component, for instance when annotators compared SVR with SVR_ENTROPY. Overall, there are still some overall predictive features: for positive influence, there are log-likelihood, standard deviation of quality score, and standard deviation of length. On the other hand, average length is a negative feature, suggesting that samples that had longer average messages were considered less interesting.

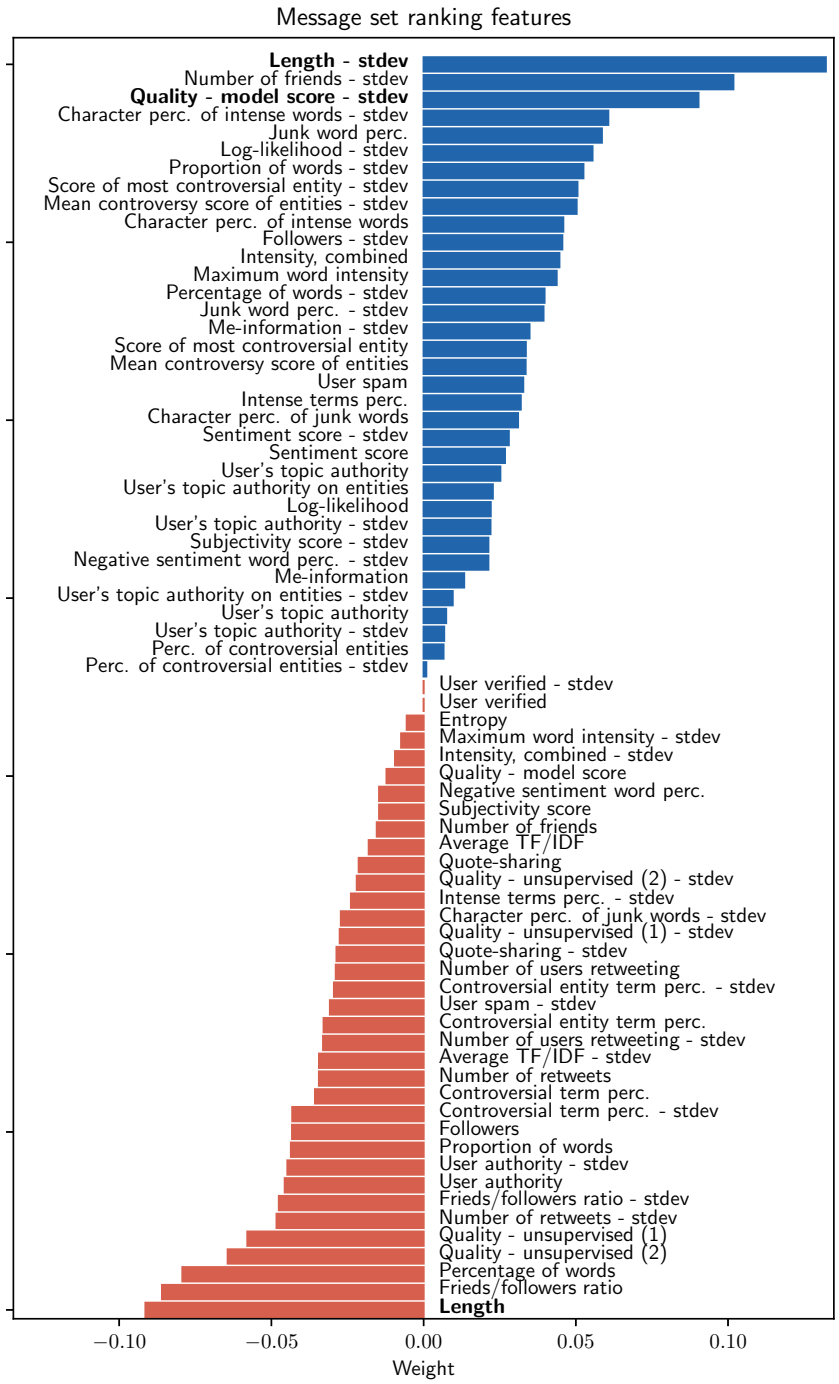


Figure 4.4: Importances of message set features as linear coefficients the preference ranking model from samples generated from all the methods using the $\mu$-$\sigma$-$H$ features. Bold names represent important features, chosen by recursive feature elimination.

## 4.2 Informal Multilingual Multi-Domain Sentiment Analysis

Sentiment analysis is a natural language processing task which aims to predict the polarity (usually denoted as positive, negative or neutral) of users publishing sentiment data, in which they express their opinions. The task is traditionally tackled as a classiication problem using supervised machine learning techniques. However, this approach requires additional effort in manual labeling of examples and often has difficulties in transferring to other domains.

One way to ameliorate this problem is to construct a lexicon of sentiment-bearing words, constructed from a wide variety of domains. While some sentiment-bearing cues are contextual, having different polarities in different contexts, the majority of words have unambiguous polarity. While this is a compromise, research shows that lexicon-based approaches can be an adequate solution if no training data is available. In practice, sentiment dictionaries or lexicons are lexical resources, which contain word associations with particular sentiments. While a number of sentiment lexicons exist for English [67], [156], the representation of sentiment resources in other languages is not as developed. The first problem this section focuses on is integrating external knowledge in the form of general-purpose sentiment lexicons.

The second problem this section focuses on is detecting sentiment in specific domains, such as social media. Besides being domain-specific, it can also be grammatically less correct and contain other properties, such as mentions of other people hash-tags, smileys and URL, as opposed to traditional movie and product review data sets.

This section explores various combinations of methods that can be used to incorporate out-of-domain training data, combined with lexicons in order to train a domain-specific sentiment classifier.

We also explore the integration of multiple data representations for a specific task of text classification. This sort of approach was also successful in the case where several combination strategies were used for the task of authorship detection [157], such as feature set concatenation or majority voting of classifiers, trained on only subsets of features. While these are known general strategies, a lot of aspects of selecting sensible feature subsets are very domain-specific.

The work from this section has been published as [158].

### 4.2.1 Sentiment lexicons

SentiWordNet [67] is the most known English-language sentiment dictionary, in which each WordNet [123] synset is represented with three numerical scores – objective Obj(s), positive Pos(s) and negative Neg(s). However, SentiWordNet does not account for domain-specificity of the input textual resources. In addition to addressing the English language, this section also discusses applications of sentiment dictionaries in Spanish. For this purpose, we have used the sentiment dictionaries published by Perez-Rosas et al. [159].

Expressing sentiment and opinion varies for different domains and document types. In such a way, sentiments carried in the news are not equivalent to the sentiments from the Twitter comments. For instance, the word "turtle" is neutral in a zoological text, but in informal Twitter comment "connection slow as a turtle", "turtle" has a negative sentiment. We evaluate a method for the construction of dictionaries as domain-specific lexical resources, which contain words, part of speech tags and the relevant sentiment scores. We have chosen the topic of telecommunication services within social media as the domain of primary interest, and the corpus, used for dictionaries development, was composed of Twitter comments referring to services of telecommunication companies. Starting with a number of positive and negative seeds for different part-of-speech words (adjectives, nouns,

verbs), these sentiment dictionaries are built in English and Spanish languages. As discussed in [160], there are a number of approaches to develop the sentiment dictionary. In our research on developing sentiment dictionaries we were following that line of work, where the authors suggested a four-step methodology for creating a domain-specific sentiment lexicon. We have modified the methodology in order to generalize it to other languages and provide sentiments for different parts of speech. We have created dictionaries not only in English, but also in Spanish. Our dictionaries were built not only for adjectives as done in [160], but also for nouns and verbs. For the English dictionary, we have additionally provided several extra features, such as the number of positive links and the number of negative links for a particular word. The English sentiment dictionary for the Telecommunication domain is composed of around 2000 adjectives, 1700 verbs and 8000 nouns, while the Spanish counterpart contains around 650 adjectives, 2000 verbs and 4100 nouns.

### 4.2.2 Feature construction

We have used different feature sources to represent individual opinion data points. In news and review datasets, every data point is a sentence, while in social media datasets, every data point is a single micro-blog post. We pre-process the textual contents by replacing URLs, numerical expressions and the names of opinions' targets with respective placeholders. We then tokenize this text, lower-case and normalize characters onto an ASCII representation due to inconsistent use of diacritics in the Spanish dataset, filter for stop-words, and weigh the terms using TF-IDF weights. The words were stemmed using the Snowball stemmer for English and Spanish [161]. The punctuation is preserved.

To accommodate social media, we have also used other text-derived features that can carry sentiment signal in informal settings, as commonly done in representation of social media text. We refer to these as **surface features** (shortened as SURF):

- count of fully capitalized words
- count of question-indicating words
- count of words that start with a capital letter
- count of repeated exclamation marks
- count of repeated same vowel
- count of repeated same character
- proportion of capital letters
- proportion of vowels
- count of negation words
- count of contrast words
- count of positive emoticons
- count of negative emoticons
- count of punctuation
- count of profanity words

We use lexicons in the form of features, where every word has assigned one or more scores. For instance, our dictionaries, described in Section 3, as well as SenticNet [162], provide a single real value in the range from -1 to 1, representing the scale from negative to positive. For these lexicons, we generate the sum of sentiment scores and the sum of absolute values of sentiment scores for every part of speech tag, as well as in total. SentiWordNet scores are represented as a triple of positive, negative and objective scores, having a total sum of 1.0. Following up on best practices, we have used a similar feature construction process as in [163], generating the following features:

- Sum of all positive sentiments of all words.
- Sum of all negative sentiment of all words.
- Total objective sentiment of all words (where $obj = 1.0 - (pos + neg)$).
- Ratio of total positive to negative scores for all words.

Besides providing total sums, we also generate these features for nouns, verbs, adjectives and adverbs separately. For Spanish, we have used the UNT sentiment lexicon [159]. Since each entry is labeled only as either positive or negative, we use the count of detected positive words and count of detected negative words as features.

### 4.2.3 Models

The data is composed of three main modalities: bag-of-words features, lexicon features, and surface features. In order to take differing distributions, dimensionality and sparsity properties into account, we use two different approaches: either concatenating the features into a single features space, or using different models for each set of features. While this situation has been solved by extending the Naïve Bayes classifier with pooling multinomials [68], we chose to implement it with a two-step model. We experiment with different feature combination approaches that are better suited for integration of background knowledge and other learning algorithms.

#### 4.2.3.1 Feature combination

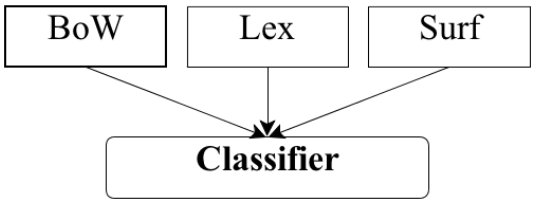


Figure 4.5: Feature concatenation diagram.

We therefore compare three feature combination approaches and a baseline, illustrated in Figures 4.5, 4.6, 4.7 and 4.8. The concatenating model simply stacks all feature spaces together and performs learning on the joint feature space. While this approach is simple, it is sensitive to different feature distributions. Therefore, we preemptively scale the features, so every feature has a standard deviation of 1.0. We do not standardize the mean, since the features themselves may be sparse, and complete standardization would turn an otherwise sparse dataset into a dense one, which is undesirable for computational

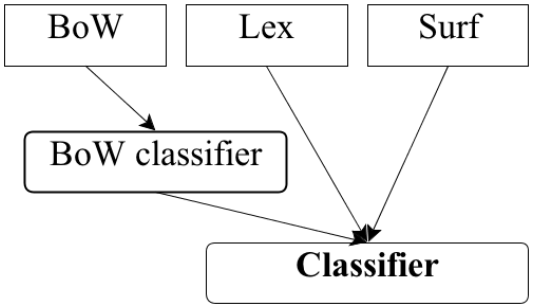


Figure 4.6: Separate bag-of-words model, denoted as "Words and features".

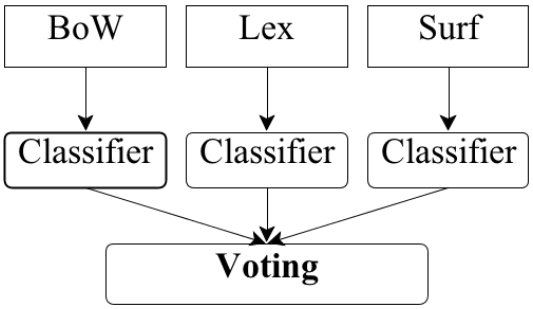


Figure 4.7: Separate model for every feature set, aggregated by voting.

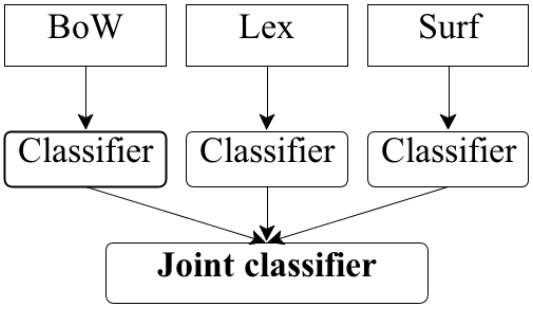


Figure 4.8: Meta-classifier, using class probabilities from the inner classifier predictions as its features.

performance reasons. The concatenation approach from Figure 4.5 is considered as the baseline. The second approach, as shown in Figure 4.6, is using a separate learning model for the bag-of-words feature set, and feeding the output of that model as features into the final classifier, together with the less sparse lexicon and surface features, in order to 'compress' the bag-of-words signal. Our third approach is related to the well-known attribute bagging [164] meta-learning strategy, with the crucial difference that the feature sub-sets are already defined in advance via domain knowledge. Our fourth approach extends the voting by employing a separate classifier model that operates on the output of the output probabilities of the inner models, in order to minimize bias of individual feature sets. We experiment by varying the training algorithm used: For the approaches using multiple models, we use the same algorithm for all the models.

All in all, we evaluate four feature set combination strategies, corresponding to Figures 4.5, 4.6, 4.7 and 4.8:

- Concatenation (Concat)
- Two-layer words and features (W+F)
- Voting model (Voting)
- Meta-classifier (Meta)

### 4.2.4 Experiments

Furthermore, we focus our experiment onto performance on our target datasets. We use the following datasets:

- Pang-Lee review dataset (PangLee), English [65], consisting of movie reviews, gathered from IMDB.
- JRC news dataset (JRC-en), English [165], consisting of statements from news articles on the topic of global politics.
- JRC news dataset, translated to Spanish using Microsoft Translator (JRC-es)
- RenderEN, English, 134 Twitter posts about a telecommunications provider (48 positive, 84 negative)
- RenderES, Spanish, 891 Twitter posts about a telecommunications provider (388 positive, 445 negative, 58 objective)

Besides our lexicons introduced in section 3 (denoted "RenderLex" and "RenderLexLinks"), we also evaluate the performance of using the Spanish lexicons [159] (denoted FullUNT and MedUNT for the full and medium variant respectively), as well as SenticNet [162] and SentiWordNet [67] for English. The label "Lex" indicates the usage of all lexicons. Our key indicators are performance metrics on RenderEN and RenderES, as they represent our use case. We perform experimental evaluation for all of these datasets on various combinations of classifiers and features construction schemes. The experiments cover various learning algorithms, as well as different modeling pipelines. We explore various combinations of feature sets: surface, bag-of-words, lexicons, as well as performance contributions of individual lexicons. The first evaluation deals with observing the applicability of various sentiment lexicons, as described in Subsection 4.2.1. First, we evaluate the lexicons in isolation, followed by a combination of lexicons together with surface features. We train a $L_1$-regularized logistic regression classifier on lexicon features. The performance is measured as averaged $F_1$-score in a 10-fold cross-validation setting.

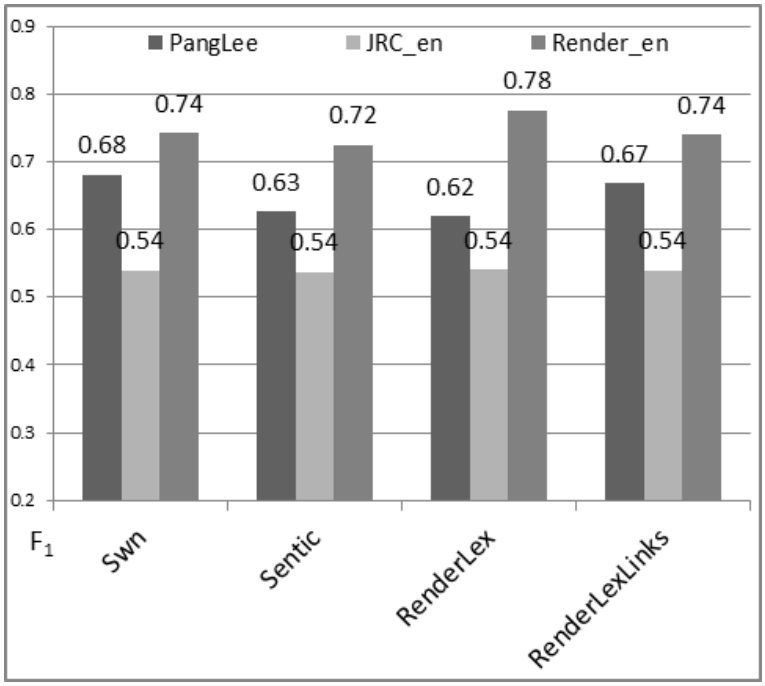


Figure 4.9: Sentiment $F_1$ scores with various sentiment lexicons for English.

Figure 4.9 shows the results, obtained performing sentiment classification on the basis of sentiment lexicon features alone. We observe that performance across the news dataset is constant, since the expression of sentiment in news does not correspond directly to sentiment meaning of individual words, but more to the domain-specific political statements. For the social media dataset, we observe improved performance when using a telecommunication domain-specific lexicon, compared to using a general domain sentiment lexicon.

While Figure 4.10 confirms the same behavior for news, the benefit of using lexicons is much lower in Spanish social media content. Given these results, we establish that a custom-built lexicon can give better results than a general purpose one. To continue, we evaluate various feature combination techniques on different learning algorithms.

Figure 4.11 displays the performance across different feature combination approaches across all datasets. Looking into individual models, we observe that the W+F model, having the bag-of-words feature set on a separate layer, consistently works best for the purpose of combining all three feature sets and masking the differences in the distribution of their features. While the W+F model consistently outperforms concatenation by a small but statistically significant margin, the Voting or Meta-classifier model only outperform concatenation on some occasions, and perform worse on the news dataset in both languages. We report the results on scenarios where LR was used as the learning algorithm on the Meta and Voting models due to the fact that they obtain a comparable performance.

Figure 4.12 shows the results on English reviews, news, and social media. On reviews, none of the additions significantly beat the bag-of-words baselines on reviews. On news, while adding SentiWordNet marginally improves the performance from 0.67 to 0.68, surface features do not give any improvement, mostly due to the formal language used in reporting, which leads to the fact that the text is written without informal cues. However, results on the Render_en social media dataset, demonstrate the performance improvements in combining all three feature sets in a two-layer model. The best performing model is able to obtain a $F_1$ score of 0.87. While the dataset is small, this demonstrates the feasibility

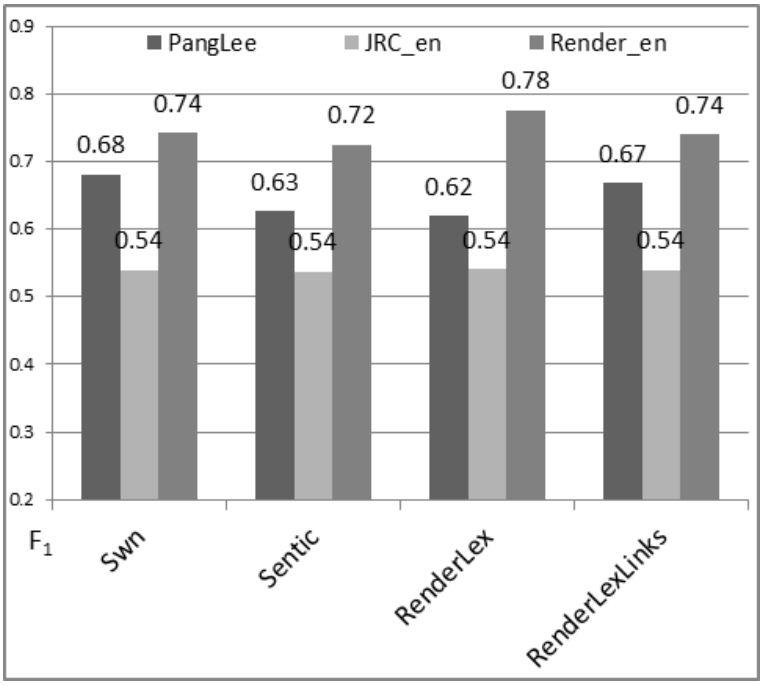


Figure 4.10: Sentiment $F_1$ scores with various sentiment lexicons for Spanish.

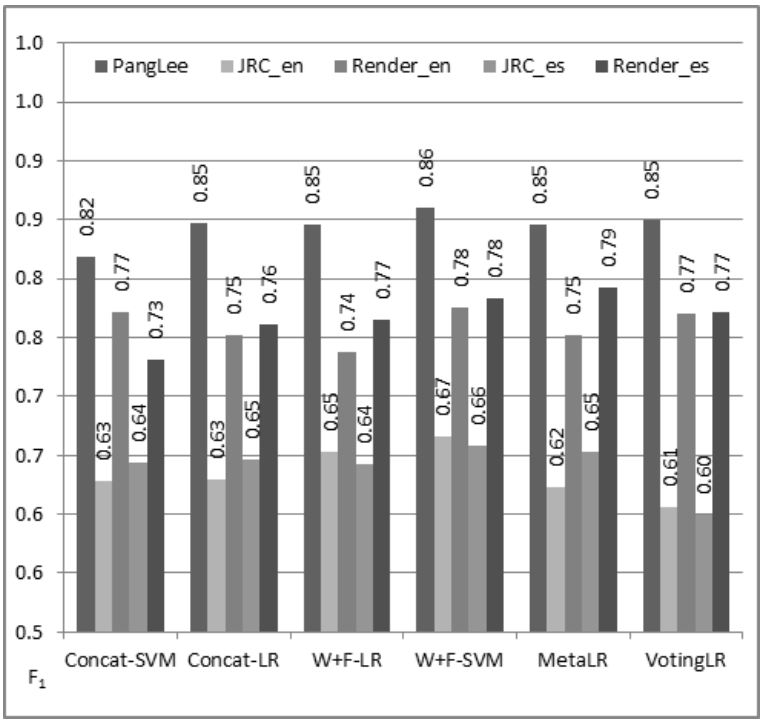


Figure 4.11: $F_1$ scores with various feature combination approaches across both languages and two learning approaches.

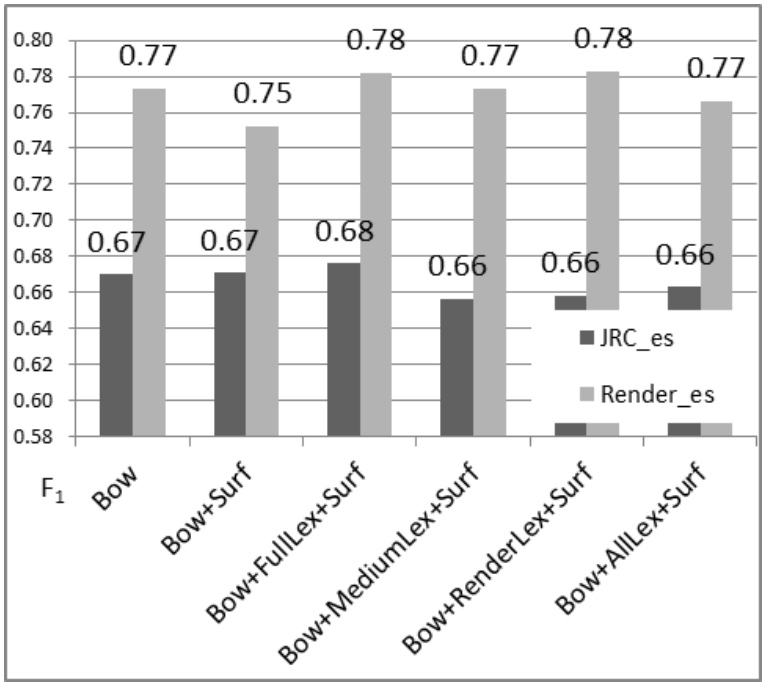


Figure 4.12: Using various feature sets on English datasets, using W+F-SVM as the model.

of using generalized external knowledge and surface features in a social media setting, especially with insufficient training data.

Figure 4.13 shows the results on both Spanish datasets when combining different feature sets in a W+F setting and a SVM model. We observe that on the news dataset, adding the Full UNT Lexicon slightly improves the $F_1$ score, while surface features alone do not give any improvement. On Render-ES, the variant combining all additions and running on a two-layer SVM model improves over the bag-of-words model by a small margin, resulting in an $F_1$ score of 0.78. Looking at the usage of various lexicons alone, it shows that the lexicons themselves only slightly improve over the surface features. In many cases, the difference is not significant, although we observe that the domain-specific lexicon RenLex does not improve over a general domain lexicon neither in news nor in social media.

### 4.2.5 Discussion

The obtained results confirm that social media content is the domain which benefits from external knowledge. Topic-specific lexicons can bring some minor improvement over general purpose lexicons, but the best-performing approaches use a combination of bag-of-words and lexicons training data. We reported improvement on two English datasets, especially on social media, which benefited significantly from preprocessing, surface features, as well as lexicons.

Moreover, having a two-layer model brings the most consistent performance across all domains and languages. In terms of comparison against state-of-the art studies, the best result on the Pang and Lee datasets scores at 0.90 F1, while ours was slightly lower at 0.88. However, on the news domain, our best approach even improves the performance on the JRC-EN dataset from the original authors' 0.65 to our result of 0.68 F1. On the other hand, the voting and meta-models did not show any improvement over the W+F

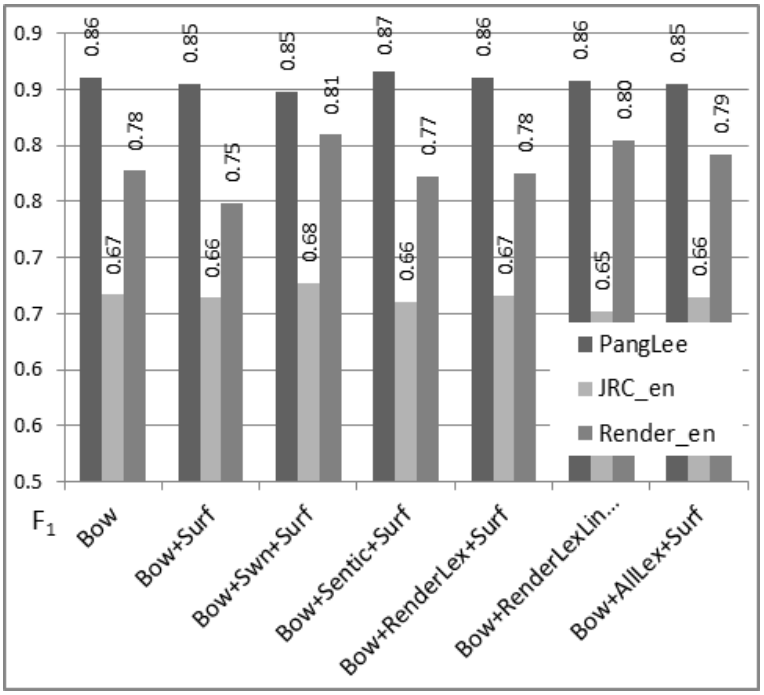


Figure 4.13: Sentiment $F_1$ scores on Spanish datasets, using W+F-SVM as the model.

model, and only improved the concatenation on some datasets, while performance was even reduced on the other datasets.

The analysis of the models shows that there are major differences between domains on which features are considered important: while news and review domains benefited from lexicons, surface features were important only in social media. On the other hand, both languages exhibited similar behavior across the same domains in news. By interpreting the models trained on social media we show that, for Spanish, surface features were more important than lexicons, while the opposite was observed for English.

A design goal of this system was to build a data-efficient sentiment analysis model that could achieve usable performance on small amounts of data. While the results reported on our datasets are lower than those typically observed on movie and product review datasets, they demonstrate improvement in multiple different scenarios of the sentiment analysis task. While the ideal system would seamlessly transfer knowledge across domains with minimal loss of performance, deployment scenarios of such tools will invariably involve domain-specific data collection, at least for quality assurance purposes. In such cases, being able to train on smaller datasets can be valuable.

The implementation of this system has been integrated in the DiversiNews system [121].

# Chapter 5

# Conclusions

This thesis presented several perspectives on the challenge of integrating a new language in an intelligent system. We described the problem domain in terms of a stack, ranging from machine learning, through natural language processing, to knowledge management. In this chapter, we summarize the contributions from individual chapters of the thesis.

## 5.1 Leveraging Knowledge Bases for Cross-Linguality

We have presented a novel method for building translation models using comparable corpora for translation of document vectors in Chapter 3. In Section 3.2, we explore different properties of the problem domain, including representation learning to take advantage of monolingual training data, kernel approximation to efficiently learn complex cross-lingual relationships, and Section 3.3 poses the problem as sparse reconstruction that allows us to give sparse solutions that correspond better to real-world translation scenarios.

In Section 3.2, we proposed a cross-lingual representation learning method using comparable corpora that uses distributional representation of words combined with a kernel approximation mapping that outperforms previous baselines while keeping its favorable scalability characteristics. We build on the existing state-of-the art approaches in vector space translation and show that learning non-linear mappings in low-dimensional spaces can improve the performance compared to linear mappings. We show that high relative link density from the source language is very important for achieving good performance. While it is difficult to distinguish between individual variations of non-linear mappings in performance, they all outperform the baselines on the task of dictionary prediction. Given the results, the most efficient method among the dominating group is SIGLSQ-, having predictable learning times that are linear with regard to the number of data points. None of the actual monolingual representation learning approaches proved significantly better than the other, but still outperform the linear approaches on most occasions.

We have also demonstrated the presented approaches on the task of similarity estimation. While all approaches using word embeddings outperform baseline approaches, we do not observe comparative benefits when using non-linear mappings to estimate similarity. The results demonstrate that for the similarity estimation task, such a system trained on a comparable corpus can even perform at an equal level of performance to a machine translation system provided that the alignments in the comparable corpus are of sufficent density.

We have evaluated the feasibility of using comparable corpora for learning language representations for the purpose of cross-lingual projection. We show that given the additional constraint of having parallel corpora, we can still perform good translation in certain cases. The main criteria for this is high relative connectedness of the source language, a

good word representation model, and a scalable translation model capable of representing non-linearities between low-dimensional representations of two languages. We can also solve the issue of relative connectedness with a hub language: literature suggests that instead of relying on sparse direct links among two languages, it is better to do two mappings using a well-connected intermediate language, such as English [166].

The results show that a method for cross-lingual vector space translation designed with sparsity in mind outperforms methods that use adaptive hard thresholding as a post-processing step on most language pairs, with the exception when translating to German. While adaptive hard thresholding itself improves performance over fixed-dimension hard thresholding, the $l1$ regularization in ElasticCCA produces even better translations in most language pairs, while CCAR still maintains the best mate retrieval performance. Results also demonstrate that some languages are easier to translate to than to translate from, for instance in Spanish.

Future work in this direction lies in use of comparable corpora for training machine translation, which traditionally uses more strictly aligned parallel corpora. The sparsity constraints in the proposed approach have the nice property that they implicitly act as disambiguators, important for reducing the search space of word translations. Other avenues of work in this direction can either be towards applying these approaches on other tasks involving natural language processing across multiple languages, such as cross-lingual textual entailment, as well as tracking news articles across languages. On the other hand, this research opens up new possibilities in using weakly aligned corpora for machine translation in the face of scarce parallel resources [167]. A possibility for further work also lies in observing the cross-lingual word correlation characteristics that could explain the anomalies when translating to and from certain languages. Another potential use case enabled by solving this problem is cross-lingual document indexing: describing a single document with a small selection of terms can potentially result in efficient and accurate inverse indices.

## 5.2 Named Entity Disambiguation

We have presented a novel method for entity resolution in texts that is based on a notion of relatedness among entities. We present several different and complementary approaches of representing relatedness and show their respective contributions to the quality of the result. In Section 3.4 we propose a framework for collective resolution of in-text entities on the basis of different notions of relatedness. As examples for this, we used three different relatedness estimation methods, each appropriate for a particular type of background knowledge. Among these methods we present and evaluate a novel method for determining relatedness based on commonness of ontological relations between two entity types and compare it to a supervised co-occurrence based approach and an approach using content similarity as relatedness. We confirm the previous related research that using collective resolution improves resolution precision, and demonstrate this on various relatedness measures. Further improvement could be obtained by the use of machine learning on other segments of the problem, such as a means of determining the importance of relations rather than calculating their selectivity. A possible application is also in determining the significance of individual relevance estimates in the last step of calculating the total assessment before making the disambiguation decisions. The proposed solution capable of entity resolution from text is an important part of the knowledge extraction. The next level of this scenario would in addition to in-text entities, also identify the relations that occur between them. These newly identified relations between entities can be a basis for constructing new RDF statements, further building our ontology, thus closing the loop where we can use existing knowledge to obtain even more knowledge. Using this technology can also be

useful for other purposes. On the other hand, we barely touch the possibilities of using globally identified data approaches, opening way for better data integration, visualization and using annotated documents to enable semantic search. We expect that the proposed semantic article enrichment method to yield even more improvement on tasks that depend on the added semantic information, such as document summarization, triple extraction and recommendation systems. What all of those use cases have in common is dependence on a high quality output of the entity resolution phase. For example, in Section 3.4.6, we describes use cases of the named entity disambiguation approaches and the standardization of the use case.

## 5.3 Automatic Selection of Interesting Social Media Responses to News

Third, we have presented a novel method for summarizing streams of responses to news in social media. We show the performance of the proposed algorithm on several real-life scenarios, as well as use it to demonstrate interesting responses, shedding light on the subjectivity around the selection process and the perception of interestingness. In Section 4.1, we proposed an optimization-driven method to solve the social message selection problem for selecting the most interesting messages posted in response to an online news article. Our method first learned the utility of individual responses in isolation and then produced a diverse selection of interesting responses by maximizing an objective function that jointly models the responses' utility scores and the entropy. The intuition behind our work was that an interesting selection of messages contains diverse, informative, opinionated and popular messages referring to the news article, written mostly by users that have authority on the topic. Our intuitions were embodied by a rich set of content, social and user features capturing these aspects.

We first compared five variations of our method with each other and then compared the best performing variation against two state of the art approaches in two experiments. The first experiment used a gold standard that consisted of messages rated by professional annotators, while the second used pair-wise comparative judgments obtained via crowdsourcing. In both experiments we found that our best performing method variation obtained the highest overall performance. The variations evaluated in these experiments also attempted using textual entailment instead of entropy as a diversification mechanism, but it did not demonstrate better sampling compared to the entropy-based sampling. Furthermore, we investigated how various indicators affect *interestingness* of an individual message, as well as of an entire message set. Our results based on the collected gold standard data found that for individual messages, *informativeness* is most important, albeit also most difficult to judge, followed by *opinionatedness* and *popularity*. We also found that message set *interestingness* was tied most strongly to *opinionatedness*, followed by *informativeness*.

The feature analysis section explained the influential features for opinionatedness, informativeness and interestingness, showing that all of the sources that we used had influential features within the three tasks. In order to understand interestingness even better, we developed a pairwise ranking model that takes message diversity into account. Our model is able to predict the more interesting message set with 69% accuracy, outperforming a baseline model that only chooses on the basis of average message score. While the model did not generalize well across samples, generated by different sampling methods, it did confirm the hypothesis that we can improve the ranking by adding new message set features computed as standard deviation of every message feature in the sample, which could serve as another proxy for diversity. One direction would be to focus on personalized models

that are based on the users' topical interests or their social circles [168]. Another possible direction is moving from extractive sampling to using methods based on abstractive summarization in order to rephrase the contents of the social media posts, which would allow for shorter, more concise summaries. We can also view the real-time nature of social media message streams as an incremental sampling setting, where we cannot assume that the entire collection of messages is available ahead of time. Our proposed approach is amenable to extension to online sampling, provided some criterion for changing the current selection of messages and substituting messages with new ones.

A possible improvement in the data collection method itself is identifying which messages are related to a news story in the first place: while having the post containing the URL of the story is a reliable indicator, it does not capture the entire discussion around that event. Rei, Grobelnik, and Mladenić [169] propose a supervised model that determines whether a tweet is related to some news event, with the supervision provided by an event database (Leban, Fortuna, Brank, *et al.* [170]).

## 5.4 Sentiment Analysis in Informal Multilingual Multi-Domain Content

In Section 4.2, we presented a novel method for sentiment analysis that uses meta-modeling and integration of background knowledge. We show its performance across multiple domains and languages, showing that it can provide good performance even on small training sets and benefit from the two-layer approach.

Besides obtaining training data, applying such a method on another language would involve obtaining a sentiment lexicon, a lemmatizer, and engineering surface features that may be specific to that language. Additional generalization would involve automating those steps or replacing them with an equivalent mechanism, such as using word embeddings or other representation learning techniques.

We also demonstrate the feasibility of using machine translation to obtain a training corpus in another language, showing that the performance obtained for JRC-ES was the same as in the original version - JRC-EN. Other research (Balahur, Steinberger, Kabadjov, *et al.* [165]) shows promising approaches to facilitate the knowledge transfer via lexicons using specifically tailored machine learning approaches. A possible avanue for future work is explore cross-lingual transfer learning, demonstrating approaches for training sentiment models using language resources from other languages.

## 5.5 Summary

The first part of the thesis in Chapter 3 demonstrated applications of cross-linguality when dealing with multiple languages on representations such as bag-of-words or named entities. The second part of the thesis in Chapter 4 observes the notion of language in social media as a specific class of language genres where assumptions on typical natural language processing tasks do not hold. Therefore, we explore possibilities of using non-textual cues in the written communication in social media, such as the social and temporal context of the message to better understand their content. While this task does not fall in the strict definition of cross-linguality, it shows that we need to treat different sub-genres of social media on different assumption sets. We demonstrated these notions on multiple tasks, such as sentiment analysis and predicting interesting message sets.

The shared challenge among both multilingual content and language in social media is that for both, we need to manage a new set of assumptions and properties of the data when

we wish to interact with them on a higher level. While training a named entity extractor on a new language requires engineering new feature sets and training corpora, the same applies to training a sentiment analysis model on a new genre of social media. From the engineering perspective, any approach that reduces this incremental cost is welcome.

We have presented multiple approaches that allow us to abstract a language when we interact with the content. We have used a notion of cross-linguality, a strategy of handling multi-lingual scenarios without employing machine translation pipelines. Furthermore, we demonstrated the possibilities and weaknesses of individual components and what kind of abstractions they let us use when dealing with multilingual content. Cross-linguality gives us the possibility to treat multilingual content on a shared representation. The representation can be either simple, like bag-of-words, or higher level, like entities, relationships, or sentiments. Either way, using shared representation opens up new possibilities for interacting with multilingual content, and the approaches presented in this thesis can obtain shared representation easier when dealing with multiple languages.

# References


[1] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning: Theory and Applications*, D. E. Holmes and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 137–186, ISBN: 978-3-540-33486-6. DOI: 10.1007/3-540-33486-6_6. [Online]. Available: https://doi.org/10.1007/3-540-33486-6_6.

[2] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *IJCAI*, vol. 7, 2007, pp. 1606–1611.

[3] G. Klyne, J. Carroll, and B. McBride, *Resource description framework (rdf): Concepts and abstract syntax*, 2004.

[4] B. Fortuna and J. Shawe-Taylor, "The use of machine translation tools for cross-lingual text mining," *Learning With Multiple Views, workshop at the ICML*, 2005.

[5] T. Cassidy, H. Ji, H. Deng, J. Zheng, and J. Han, "Analysis and refinement of cross-lingual entity linking," in *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, Springer, 2012, pp. 1–12.

[6] K. Duh, A. Fujino, and M. Nagata, "Is machine translation ripe for cross-lingual sentiment classification?" In *ACL (Short Papers)*, 2011, pp. 429–433.

[7] A. Garzó, B. Daróczy, T. Kiss, D. Siklósi, and A. A. Benczúr, "Cross-lingual web spam classification," in *Proceedings of the 22nd international conference on World Wide Web companion*, International World Wide Web Conferences Steering Committee, 2013, pp. 1149–1156.

[8] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *CoRR*, vol. abs/1309.4168, 2013.

[9] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," Association for Computational Linguistics, 2014.

[10] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1393–1398.

[11] O. Levy, A. Søgaard, and Y. Goldberg, "A strong baseline for learning cross-lingual word embeddings from sentence alignments," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 765–774.

[12] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, vol. 45, 2007, p. 976.

[13] I. Vulić and M.-F. Moens, "Bilingual distributed word representations from document-aligned comparable data," *Journal of Artificial Intelligence Research*, vol. 55, pp. 953–994, 2016.

[14] A. Søgaard, Ž. Agić, H. M. Alonso, B. Plank, B. Bohnet, and A. Johannsen, "Inverted indexing for cross-lingual nlp," in *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, 2015.

[15] S. Ruder, "A survey of cross-lingual embedding models," *arXiv preprint arXiv:1706.04902*, 2017.

[16] P. Fung, "A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora," in *Machine Translation and the Information Soup*, Springer, 1998, pp. 1–17.

[17] I. Vulic and M.-F. Moens, "Cross-lingual semantic similarity of words as the similarity of their semantic word responses," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, ACL, 2013, pp. 106–116.

[18] A. Mogadala and A. Rettinger, "Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification.," in *HLT-NAACL*, 2016, pp. 692–702.

[19] M. L. Littman, S. T. Dumais, and T. K. Landauer, "Automatic cross-language information retrieval using latent semantic indexing," in *Cross-language information retrieval*, Springer, 1998, pp. 51–62.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[21] P. Sorg and P. Cimiano, "Exploiting wikipedia for cross-lingual and multilingual information retrieval," *Data & Knowledge Engineering*, vol. 74, pp. 26–45, 2012.

[22] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini, "Inferring a semantic representation of text via cross-language correlation analysis," *Advances in neural information processing systems*, vol. 15, pp. 1473–1480, 2003.

[23] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[24] J. Rupnik and B. Fortuna, "Regression canonical correlation analysis," in *Learning from Multiple Sources, NIPS Workshop*, vol. 13, 2008.

[25] J. Rupnik, A. Muhic, and P. Škraba, "Low-rank approximations for large, multilingual data," *Avaliable at http://ailab. ijs. si/primoz_ skraba/papers/nips_full. pdf*, 2011.

[26] J. Ni, G. Dinu, and R. Florian, "Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection," *arXiv preprint arXiv:1707.02483*, 2017.

[27] P. Klein, S. P. Ponzetto, and G. Glavaš, "Improving neural knowledge base completion with cross-lingual projections," Association for Computational Linguistics, 2017.

[28] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[29] M. Mancini, J. Camacho-Collados, I. Iacobacci, and R. Navigli, "Embedding words and senses together via joint knowledge-enhanced training," in *21st Conference on Computational Natural Language Learning (CoNLL 2017)*, ACM, 2017, pp. 100–111.

[30] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969.

[31] L. H. R. M. B. Niswonger, M. T. Roth, P. Schwarz, and E. Wimmers, "Transforming heterogeneous data with database middleware: Beyond integration," *Data Engineering*, vol. 22, no. 1, pp. 31–36, 1999.

[32] W. E. Winkler, "The state of record linkage and current research problems," in *Statistical Research Division, US Census Bureau*, Citeseer, 1999.

[33] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," *Information Systems*, vol. 26, no. 8, pp. 607–633, 2001.

[34] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 1, pp. 1–16, 2007.

[35] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1995, pp. 189–196.

[36] D. V. Kalashnikov and S. Mehrotra, "A probabilistic model for entity disambiguation using relationships," in *SIAM International Conference on Data Mining (SDM). Newport Beach, California*, 2005, pp. 21–23.

[37] H. Schütze, "Automatic word sense discrimination," *Computational linguistics*, vol. 24, no. 1, pp. 97–123, 1998.

[38] R. C. Bunescu and M. Pasca, "Using encyclopedic knowledge for named entity disambiguation.," in *EACL*, vol. 6, 2006, pp. 9–16.

[39] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data.," in *EMNLP-CoNLL*, vol. 7, 2007, pp. 708–716.

[40] R. Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2005, pp. 411–418.

[41] P. Singla and P. Domingos, "Entity resolution with markov logic," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*, IEEE, 2006, pp. 572–582.

[42] C. Ramakrishnan, W. H. Milnor, M. Perry, and A. P. Sheth, "Discovering informative connection subgraphs in multi-relational graphs," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 56–63, 2005.

[43] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 5, 2007.

[44] A. Lommatzsch, D. Ploch, E. W. De Luca, and S. Albayrak, "Named entity disambiguation for german news articles," *WIR 2010*,

[45] V. I. Spitkovsky and A. X. Chang, "Strong baselines for cross-lingual entity linking," *TAC 2011*, 2011.

[46] P. McNamee, J. Mayfield, D. Lawrie, D. W. Oard, and D. S. Doermann, "Cross-language entity linking.," in *IJCNLP*, 2011, pp. 255–263.

[47] H. Becker, M. Naaman, and L. Gravano, "Selecting quality Twitter content for events," in *Proceedings of the 2011 AAAI International Conference on Weblogs and Social Media*, 2011.

[48] M. De Choudhury, S. Counts, and M. Czerwinski, "Find me the right content! Diversity-based sampling of social media spaces for topic-centric search," in *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*, 2011.

[49] B. Sharifi, M. Hutton, and J. Kalita, "Experiments in microblog summarization," in *Proceedings of the 2nd IEEE International Conference on Social Computing*, 2010, pp. 49–56.

[50] F. Liu, Y. Liu, and F. Weng, "Why is "SXSW"trending? Exploring multiple text sources for Twitter topic summarization," in *Proceedings of the 11th ACL Conference on Human Language Technologies*, 2010, pp. 66–75.

[51] G. Beverungen and J. Kalita, "Evaluating methods for summarizing Twitter posts," in *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*, 2011.

[52] N. Naveed, T. Gottron, J. Kunegis, and A. Alhadi, "Bad news travel fast: a content-based analysis of interestingness on Twitter," in *Proceedings of the 3rd Annual ACM Web Science Conference*, 2011.

[53] O. Tsur and A. Rappoport, "What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012, pp. 643–652.

[54] E. Lloret and M. Palomar, "Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre," *Expert Systems with Applications*, vol. 40, no. 16, 2013.

[55] M. De Choudhury, H. Sundaram, A. John, and D. Seligmann, "What makes conversations interesting?: Themes, participants and consequences of conversations in online social media," in *Proceedings of the 18th International Conference on World Wide Web*, 2009, pp. 331–340.

[56] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*, 2008, pp. 183–194.

[57] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: characterizing emerging trends on Twitter," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 5, pp. 902–918, 2011.

[58] A.-M. Popescu and M. Pennacchiotti, ""Dancing with the Stars", NBA games, politics: an exploration of Twitter users' response to events," in *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*, 2011.

[59] Z. Luo, M. Osborne, and T. Wang, "Opinion retrieval in Twitter," in *Proceedings of the 6th AAAI International Conference on Weblogs and Social Media*, 2012.

[60] O. Dalal, S. Sengemedu, and S. Sanyal, "Multi-objective ranking of comments on web," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 419–428.

[61] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha, "Time is of the essence: improving recency ranking using Twitter data," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 331–340.

[62] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, 2012, pp. 189–198.

[63] R. Bandari, S. Asur, and B. Huberman, "The pulse of news in social media: forecasting popularity," in *Proceedings of the 6th AAAI International Conference on Weblogs and Social Media*, 2012.

[64] Z. Yang, K. Cai, J. Tang, L. Zhang, Z. Su, and J. Li, "Social context summarization," in *Proceedings of the 34th ACM International Conference on Research and Development in Information Retrieval*, 2011, pp. 255–264.

[65] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, 2002, pp. 79–86.

[66] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[67] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6, 2006, pp. 417–422.

[68] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2009, pp. 1275–1284.

[69] Y. Mejova and P. Srinivasan, "Crossing media streams with sentiment: Domain adaptation in blogs, reviews and twitter.," in *ICWSM*, 2012.

[70] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. Van der Goot, "Jrc-names: A freely available, highly multilingual named entity resource," *arXiv preprint arXiv:1309.6162*, 2013.

[71] T. Štajner and M. Grobelnik, "Story link detection with entity resolution," in *WWW 2009 Workshop on Semantic Search*, vol. 2, 2009, pp. 3–2.

[72] F. Su, B. Babych, M. Paramite, and R. Gaizauskas, "Evaluation and Elaboration of Metrics," *ACCURAT Project Deliverable D1.3*, Dec. 2011.

[73] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[74] B. Thompson, "Canonical correlation analysis," *Encyclopedia of statistics in behavioral science*, 2005.

[75] T. Štajner and D. Mladenić, "Cross-lingual document similarity estimation and dictionary generation with comparable corpora," *Knowledge and Information Systems*, pp. 1–15, 2018.

[76] L. Zhang, A. Rettinger, M. Färber, and M. Tadić, "A comparative evaluation of cross-lingual text annotation techniques," in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Springer, 2013, pp. 124–135.

[77] S. Lauly, A. Boulanger, and H. Larochelle, "Learning multilingual word representations using a bag-of-words autoencoder," *arXiv preprint arXiv:1401.1803*, 2014.

[78] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[79] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 160–167.

[80] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2177–2185.

[81] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, 2001, pp. 682–688.

[82] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random fourier features: A theoretical and empirical comparison," in *NIPS*, 2012, pp. 485–493.

[83] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Y. Lechevallier and G. Saporta, Eds., Paris, France: Springer, Aug. 2010, pp. 177–187. [Online]. Available: `http://leon.bottou.org/papers/bottou-2010`.

[84] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 3, pp. 480–492, 2012.

[85] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[86] S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer, "Automatic cross-language retrieval using latent semantic indexing," in *AAAI spring symposium on cross-language text and speech retrieval*, vol. 15, 1997, p. 21.

[87] J. C. A. Barata and M. S. Hussein, "The moore–penrose pseudoinverse: A tutorial review of the theory," *Brazilian Journal of Physics*, vol. 42, no. 1-2, pp. 146–165, 2012.

[88] S. Hellmann, J. Brekle, and S. Auer, "Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud," in *Semantic Technology*, Springer, 2013, pp. 191–206.

[89] J. Rupnik, A. Muhic, G. Leban, P. Skraba, B. Fortuna, and M. Grobelnik, "News across languages-cross-lingual document similarity and event tracking," *Journal of Artificial Intelligence Research*, vol. 55, pp. 283–316, 2016.

[90] M. L. Paramita, P. D. Clough, A. Aker, and R. J. Gaizauskas, "Correlation between similarity measures for inter-language linked wikipedia articles.," in *LREC*, 2012, pp. 790–797.

[91] A. Barrón-Cedeno, M. L. Paramita, P. Clough, and P. Rosso, "A comparison of approaches for measuring cross-lingual similarity of wikipedia articles.," in *ECIR*, 2014, pp. 424–429.

[92] E. Morin and E. Prochasson, "Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora," in *Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web*, Association for Computational linguistics, 2011, pp. 27–34.

[93] A. Laroche and P. Langlais, "Revisiting context-based projection methods for term-translation spotting in comparable corpora," in *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics, 2010, pp. 617–625.

[94] R. G. Baraniuk, "Compressive sensing [lecture notes]," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 118–121, 2007.

[95] Z. Deng, G. Hu, Z. Pan, and Y. Zhang, "Kernel sparse feature selection based on semantics in text classification," 2012.

[96] L. Yu, "Linked open data," in *A Developer's Guide to the Semantic Web*, Springer, 2011, pp. 409–466.

[97] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*, Association for Computational Linguistics, 2003, pp. 1–8.

[98] T. Štajner and D. Mladenić, "Entity resolution in texts using statistical learning and ontologies," in *Asian Semantic Web Conference*, Springer, 2009, pp. 91–104.

[99] A. McCallum, "Information extraction: Distilling structured data from unstructured text," *Queue*, vol. 3, no. 9, pp. 48–57, 2005.

[100] L. Lloyd, V. Bhagwan, D. Gruhl, and A. Tomkins, "Disambiguation of references to individuals," *IBM Technical Report*, 2005.

[101] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[102] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2005, pp. 363–370.

[103] X. Li, P. Morie, and D. Roth, "Semantic integration in text: From ambiguous names to identifiable entities," *AI magazine*, vol. 26, no. 1, p. 45, 2005.

[104] S. E. Overell, J. Magalhaes, S. M. Rüger, *et al.*, "Place disambiguation with co-occurrence models.."

[105] M.-G. Jang, S. H. Myaeng, and S. Y. Park, "Using mutual information to resolve query translation ambiguities and query term weighting," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, 1999, pp. 223–229.

[106] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[107] H. Li and N. Abe, "Word clustering and disambiguation based on co-occurrence data," in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, 1998, pp. 749–755.

[108] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[109] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," 2013.

[110] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, Springer, 2007, pp. 722–735.

[111] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007, pp. 697–706.

[112] E. Sandhaus, "The new york times annotated corpus," *Linguistic Data Consortium, Philadelphia*, vol. 6, no. 12, e26752, 2008.

[113] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang, "Modeling mention, context and entity with neural networks for entity disambiguation.," in *IJCAI*, 2015, pp. 1333–1339.

[114] A. Barrena, A. Soroa, and E. Agirre, "Alleviating poor context with background knowledge for named entity disambiguation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1903–1912.

[115] T. Štajner and D. Mladenić, "Cross-lingual named entity extraction and disambiguation," in *4th Jožef Stefan International Postgraduate School Students Conference*, 2012.

[116] A. Rettinger, L. Zhang, D. Berović, D. Merkler, M. Srebačić, and M. Tadić, "Recsa: Resource for evaluating cross-lingual semantic annotation.," in *LREC*, 2014, pp. 4000–4003.

[117] T. Štajner, T. Erjavec, and S. Krek, "Razpoznavanje imenskih entitet v slovenskem besedilu," *Slovenščina 2.0*, 2013.

[118] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, Association for Computational Linguistics, 2003, pp. 188–191.

[119] T. Štajner, D. Rusu, L. Dali, B. Fortuna, D. Mladenić, and M. Grobelnik, "A service oriented framework for natural language text enrichment," *Informatica*, vol. 34, no. 3, 2010.

[120] L. Padró, Ž. Agić, X. Carreras, B. Fortuna, E. García Cuesta, Z. Li, T. Štajner, and M. Tadić, "Language processing infrastructure in the xlike project," in *LREC 2014: Ninth International Conference on Language Resources and Evaluation: Reykjavik, Islàndia: May, 26-31, 2014: proceedings*, European Language Resources Association (ELRA), 2014, pp. 3811–3816.

[121] M. Trampuš, F. Fuart, D. Pighin, T. Štajner, J. Berčič, B. Novak, D. Rusu, L. Stopar, and M. Grobelnik, "Diversinews: Surfacing diversity in online news," *AI Magazine*, vol. 36, no. 4, pp. 87–104, 2015.

[122] D. Filip, S. McCance, D. Lewis, C. Lieske, A. Lommel, J. Kosek, F. Sasaki, and Y. Savourel, *Internationalization tag set (its) version 2.0*, 2012. [Online]. Available: `http://www.w3.org/TR/its20/`.

[123] C. Fellbaum, *WordNet*. Wiley Online Library, 1999.

[124] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer, "Integrating nlp using linked data," in *The Semantic Web – ISWC 2013*, ser. Lecture Notes in Computer Science, H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, Eds., vol. 8219, Springer Berlin Heidelberg, 2013, pp. 98–113, ISBN: 978-3-642-41337-7. DOI: `10.1007/978-3-642-41338-4_7`. [Online]. Available: `http://dx.doi.org/10.1007/978-3-642-41338-4_7`.

[125] T. Consortium, L. Burnard, S. Bauman, *et al.*, *TEI P5: Guidelines for electronic text encoding and interchange*. TEI Consortium, 2008.

[126] M. Grobelnik and D. Mladenić, "Simple classification into large topic ontology of web documents," *CIT. Journal of computing and information technology*, vol. 13, no. 4, pp. 279–285, 2005.

[127] M. Pinnis, N. Ljubešic, D. Stefanescu, I. Skadina, M. Tadić, and T. Gornostay, "Term extraction, tagging, and mapping tools for under-resourced languages," in *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, 2012, pp. 20–21.

[128] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes, "Automatic selection of social media responses to news," in *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 50–58.

[129] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from Twitter," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 1873–1876.

[130] F. Zanzotto, M. Pennacchiotti, and K. Tsioutsiouliklis, "Linguistic redundancy in Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 659–669.

[131] E. Bakshy, J. Hofman, W. Mason, and D. Watts, "Everyone's an influencer: quantifying influence on Twitter," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 2011, pp. 65–74.

[132] J. An, D. Quercia, and J. Crowcroft, "Why individuals seek diverse opinions (or why they don't)," in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 15–18.

[133] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[134] M. Negri, M. Kouylekov, B. Magnini, Y. Mehdad, and E. Cabrio, "Towards extensible textual entailment engines: the EDITS package," *Proceedings of the 11th International Conference of the Italian Association for Artificial Intelligence*, pp. 314–323, 2009.

[135] M. Kouylekov, Y. Mehdad, M. Negri, and E. Cabrio, "FBK participation in RTE6: main and KBP validation task," in *Proceedings of the 6th Recognizing Textual Entailment Challenge*, 2010.

[136] M. Naaman, J. Boase, and C. Lai, "Is it really about me? Message content in social awareness streams," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010, pp. 189–192.

[137] S. Ye and S. Wu, "Measuring message propagation and social influence on Twitter.com," in *Proceedings of the 2nd International Conference on Social Informatics*, 2010, pp. 216–231.

[138] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in Twitter: the million follower fallacy," in *Proceedings of the 2010 AAAI International Conference on Weblogs and Social Media*, 2010.

[139] D. Paranjpe, "Learning document aboutness from implicit user feedback and document structure," in *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, 2009, pp. 365–374.

[140] C. Long, M. Huang, X. Zhu, and M. Li, "A new approach for multi-document update summarization," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 739–749, 2010.

[141] K. Ganesan, C. Zhai, and E. Viegas, "Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 869–878.

[142] Y. Chang, X. Wang, Q. Mei, and Y. Liu, "Towards twitter context summarization with user influence models," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, 2013, pp. 527–536.

[143] J. Cohen, "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[144] C.-Y. Lin, "ROUGE: a package for automatic evaluation of summaries," in *Proceedings of the 2004 ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[145] D. Das and A. F. Martins, "A survey on automatic text summarization," *Literature Survey for the Language and Statistics II course at CMU*, vol. 4, pp. 192–195, 2007.

[146] U. Gadiraju, "It's getting crowded!: Improving the effectiveness of microtask crowdsourcing," PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover, 2017.

[147] T. Cuzzillo, *Real-World Active Learning, Applications and Strategies for Human-in-the-loop Machine Learning*. O'Reilly, 2015, p. 9.

[148] J. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[149] S. Gouws, D. Metzler, C. Cai, and E. Hovy, "Contextual bearing on linguistic variation in social media," in *Proceedings of the 12th ACM Conference on Human Language Technologies*, 2011, p. 20.

[150] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[151] M. T. Ribeiro, S. Singh, and C. Guestrin, ""WHY SHOULD I TRUST YOU?": EXPLAINING THE PREDICTIONS OF ANY CLASSIFIER," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[152] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Advances in Advances in Large Margin Classifiers*, pp. 115–132, 1999.

[153] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.

[154] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, 2011.

[155] O. Chapelle and S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, vol. 13, no. 3, pp. 201–215, 2010.

[156] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Computational Linguistics and Intelligent Text Processing*, Springer, 2005, pp. 486–497.

[157] A. Kaster, S. Siersdorfer, and G. Weikum, "Combining text and linguistic document representations for authorship attribution," in *SIGIR workshop: Stylistic Analysis of Text For Information Access*, 2005.

[158] T. Štajner, I. Novalija, and D. Mladenić, "Informal multilingual multi-domain sentiment analysis," *Informatica*, vol. 37, no. 4, 2013.

[159] V. Perez-Rosas, C. Banea, and R. Mihalcea, "Learning sentiment lexicons in spanish.," in *LREC*, 2012, pp. 3077–3081.

[160] A. Bizău, D. Rusu, and D. Mladenić, "Expressing opinion diversity," *DiversiWeb 2011*, p. 5, 2011.

[161] M. F. Porter, *Snowball: A language for stemming algorithms*, 2001.

[162] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis.," in *FLAIRS Conference*, 2012, pp. 202–207.

[163] B. Ohana and B. Tierney, "Sentiment classification of reviews using sentiwordnet," in *9th. IT & T Conference*, 2009, p. 13.

[164] R. Bryll, R. Gutierrez-Osuna, and F. Quek, "Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets," *Pattern recognition*, vol. 36, no. 6, pp. 1291–1302, 2003.

[165] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *LREC*, 2010.

[166] J. Rupnik, A. Muhic, and P. Škraba, "Cross-lingual document retrieval through hub languages," 2012.

[167] I. Skadiņa, A. Aker, N. Mastropavlos, F. Su, D. Tufis, M. Verlic, A. Vasiļjevs, B. Babych, P. Clough, R. Gaizauskas, *et al.*, "Collecting and using comparable corpora for statistical machine translation," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012.

[168] M. Bernstein, L. Hong, S. Kairam, H. Chi, and B. Suh, "A torrent of tweets: Managing information overload in online social streams," in *In Proceedings of the 2010 CHI Workshop on Microblogging: What and How Can We Learn From It?*, 2010.

[169] L. Rei, M. Grobelnik, and D. Mladenić, "Event detection in twitter with an event knowledge base," in *Izkopavanje znanja in podatkovna skladišča (SiKDD 2015): zbornik 18. mednarodne multikonference Informacijska družba - IS 2015, 5. oktober 2015, [Ljubljana, Slovenia]: zvezek E.* 2015, pp. 33–36. [Online]. Available: `http://dirros.openscience.si/IzpisGradiva.php?lang=slv&id=884`.

[170] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," in *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 2014, pp. 107–110.

# Bibliography

## Publications Related to the Thesis

### Journal Articles

T. Štajner and D. Mladenić, "Cross-lingual document similarity estimation and dictionary generation with comparable corpora," *Knowledge and Information Systems*, pp. 1–15, 2018.

T. Štajner, T. Erjavec, and S. Krek, "Razpoznavanje imenskih entitet v slovenskem besedilu," *Slovenščina 2.0*, 2013.

T. Štajner, D. Rusu, L. Dali, B. Fortuna, D. Mladenić, and M. Grobelnik, "A service oriented framework for natural language text enrichment," *Informatica*, vol. 34, no. 3, 2010.

T. Štajner, I. Novalija, and D. Mladenić, "Informal multilingual multi-domain sentiment analysis," *Informatica*, vol. 37, no. 4, 2013.

### Conference Papers

T. Štajner and D. Mladenić, "Entity resolution in texts using statistical learning and ontologies," in *Asian Semantic Web Conference*, Springer, 2009, pp. 91–104.

——, "Cross-lingual named entity extraction and disambiguation," in *4th Jožef Stefan International Postgraduate School Students Conference*, 2012.

T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes, "Automatic selection of social media responses to news," in *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 50–58.

## Other Publications

T. Štajner and M. Grobelnik, "Story link detection with entity resolution," in *WWW 2009 Workshop on Semantic Search*, vol. 2, 2009, pp. 3–2.

L. Padró, Ž. Agić, X. Carreras, B. Fortuna, E. García Cuesta, Z. Li, T. Štajner, and M. Tadić, "Language processing infrastructure in the xlike project," in *LREC 2014: Ninth International Conference on Language Resources and Evaluation: Reykjavik, Islàndia: May, 26-31, 2014: proceedings*, European Language Resources Association (ELRA), 2014, pp. 3811–3816.

M. Trampuš, F. Fuart, D. Pighin, T. Štajner, J. Berčič, B. Novak, D. Rusu, L. Stopar, and M. Grobelnik, "Diversinews: Surfacing diversity in online news," *AI Magazine*, vol. 36, no. 4, pp. 87–104, 2015.

# Biography

The author of this thesis is a Ph. D. candidate at Jožef Stefan International Post-graduate School and a former researcher at Artificial Intelligence Laboratory at Jožef Stefan Institute in Ljubljana. He received his diploma degree from the Faculty of Computer and Information Science at University of Ljubljana, with the thesis "Entity Resolution in Texts Using Machine Learning and Background Knowledge". His main research interests are in applying machine learning to natural language processing, speech recognition, with application on navigation and mobility use-cases. He has participated in several FP7 research projects, co-founded a start-up and worked in multiple other start-ups. He currently works at HERE Technologies in Berlin as a Lead Data Scientist.